Eötvös Loránd Tudományegyetem
Társadalomtudományi Kar
ALAPKÉPZÉS

Társadalmi torzítások a gépi tanulásban
Esettanulmány a Google Fordítóról

Konzulens:
Németh Renáta

Készítette:
Farkas Anna
RRK6PV
szociológia szak

2020. május

Köszönetnyilvánítás

Szeretném megköszönni konzulensemnek, Németh Renátának, aki ötleteivel és tanácsaival folyamatosan segítette a munkámat.

Szeretnék köszönetet mondani az Insipira Group kutatócégnek, akik a szakdolgozatomhoz készült kérdőív lekérdezésében nyújtottak segítséget.

Továbbá köszönettel tartozom családomnak a támogatásukért.

Tartalomjegyzék

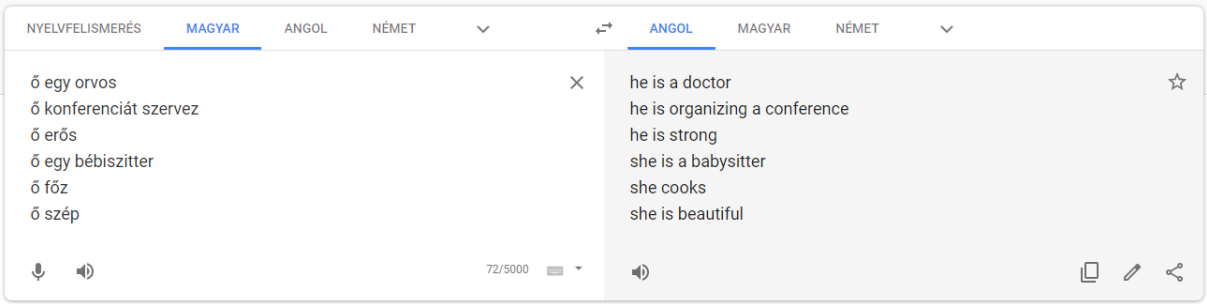
Absztrakt

A szakdolgozat témája a gépi tanuló algoritmusokban előforduló társadalmi torzítások kialakulása, felderítése, mérése és kijavításának lehetősége. A közelmúltban több kutatás is készült azzal kapcsolatban, hogy a gépi tanuló algoritmusok hajlamosak megismételni vagy felerősíteni a valós társadalmi különbségeket, előítéleteket, sztereotípiákat és így olyan döntéseket hozhatnak, ami hátrányban részesít társadalmi csoportokat. Ennek oka a gépi tanuláson alapuló algoritmusok működéséhez és tanulásához szükséges adatok gyűjtésekor és feldolgozásakor keletkező társadalmi torzítások. Ezt a jelenséget a szakdolgozat egy Google Fordítóról készült esettanulmányon keresztül mutatja be, amely a gépi fordításban megjelenő nemi torzítást vizsgálja foglalkozások és hozzájuk kapcsolt jelzők magyar-angol fordításánál.

1. Bevezetés

Az életünket alakító döntésekben egyre gyakrabban kapnak teret úgynevezett gépi tanuláson alapuló algoritmusok (Ságvári 2017). Ilyen algoritmusok határozzák meg a keresési találatainkat; szabályozzák, hogy milyen hirdetéseket látunk a közösségi oldalakon; döntenek hitelkérelmekről; diagnosztizálnak betegségeket és számos más területen is alkalmazzák azokat (Burrell 2016; Goodman–Flaxman 2016; Sandvig és társai 2014; Ságvári 2017). Egyszóval egyéneket, csoportokat és a társadalom egészét befolyásoló döntésekben kapnak teret a gépi tanuló algoritmusok. Azt gondoljuk és azt várjuk, hogy ezek az algoritmusok objektívek legyenek, ne jellemezze őket a részrehajlás, mégis, egyre több tanulmány (Barocas–Selbst 2016; Chen és társai 2018; Ságvári 2017) és cikk (Angwin és társai 207; Dastin 2018; Schwarm 2018) foglalkozik azzal, hogyan járulnak hozzá az algoritmusok diszkriminatív, részrehajló döntések születéséhez. Vagyis egyre több kutatás foglalkozik azzal a jelenséggel, hogy a gépi tanuláson alapuló rendszerek leképezik vagy felerősítik a társadalomban meglévő nemi, faji, vallási vagy kor alapú diszkriminációt. Szakdolgozatomban ezzel a jelenséggel, a gépi tanulásban előforduló társadalmi torzításokkal és ennek diszkriminatív hatásával foglalkozom. Azzal, hogy miért torzítanak az algoritmusok, miért válnak diszkriminatívvá és ezt hogyan lehet vizsgálni. Miután áttekintettem, hogy hogyan és miért vezethetnek az alapvetően racionálisan működő, objektívnek tartott algoritmusok társadalmi igazságtalansághoz, a jelenség gyakorlati vizsgálatát egy esettanulmányon keresztül mutatom be, ami a Google Fordítóról, egy gépi tanuláson alapuló online fordítóprogramról készült.

Esettanulmányomban azt vizsgáltam, hogy hogyan jelenik meg a nemi torzítás (*gender bias*) a Google Fordítóban foglalkozások magyar-angol fordításánál. A magyar egy nemsemleges (*gender neutral*) nyelv, ami azt jelenti, hogy az egyes szám harmadik személyt jelölő névmás („ő") nem különbözik a nemek esetében, szemben az angol nyelvvel, ahol az egyes szám harmadik személyt jelölő névmás lehet hímnemű („he"), nőnemű („she") és semleges („it"). A fordítóprogramoknál a nemsemleges nyelvekről nemeket megkülönböztető (*gender-based*) nyelvekre történő fordítás esetében az eredetileg semleges személyes névmás a fordításban nem csak semleges, hanem hímnemű és nőnemű is lehet. Korábban ez így volt a Google Fordító esetében is. Ha beírtunk olyan mondatokat a fordítóba, mint „ő egy orvos", „ő konferenciát szervez" vagy „ő erős", ahol a kontextusból nem adódik sem az, hogy az alany férfi, sem az, hogy nő lenne, úgy fordította le azokat angolra, hogy „he is a doctor", „he is organizing a conference" és „he is strong". Tehát az angol fordításban hímnemű személyes névmást használt. Az olyan mondatokat, mint „ő egy bébiszitter", „ő főz" és „ő szép" pedig úgy fordította, hogy „she is a babysitter", „she cooks" és „she is beautiful", ezekben az esetekben az angol fordításban nőnemű személyes névmást használt. Ezt mutatja be az 1. Ábra[1].

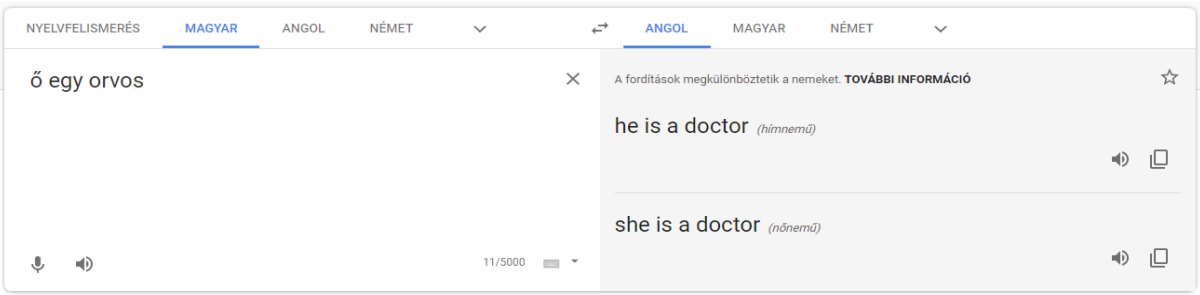


1. Ábra

Tehát azok a férfiakkal és nőkkel szembeni sztereotípiák, amik a társadalomban élnek, legyen szó a foglalkozásoktól kezdve olyan tulajdonságokig, mint „szép" és „erős", a fordítóprogramban is megjelent. Azért, mert minden szöveget egyféleképpen fordított le, még akkor is, ha annak semleges, férfi és női megfelelője is volt a célnyelvben, mint az „ő" angol

[1] Forrás: Google Fordító [Képernyőkép] (https://translate.google.hu/?hl=hu&tab=rT#view=home&op=translate&sl=hu&tl=en&text=%C5%91%20egy%20orvos%0A%C5%91%20konferenci%C3%A1t%20szervez%0A%C5%91%20er%C5%91s%0A%C5%91%20egy%20b%C3%A9biszitter%0A%C5%91%20f%C5%91z%0A%C5%91%20sz%C3%A9p) (Utolsó megtekintés: 2020. április 23.)

fordításánál az „it”, a „he” és a „she”. Az ebből adódó nemi torzítás bármely nemsemleges nyelvről (pl. magyarról) a nemeket megkülönböztető nyelvekre (pl. angolra) történő fordításnál megjelenhetett.



2. Ábra

A Google Fordító 2018-tól kezdve bizonyos nyelveknél (pl. török-angol fordításoknál) elérhetővé tette a nemsemleges szavak esetében mind a hímnemű, mind a nőnemű fordítást. Az „o bir doktor” („ő egy orvos”) török mondatot ettől kezdve kétféleképpen fordítja le angolra a program: „she is a doctor” és „he is a doctor” (Kuczmarski 2018). Az ilyen esetekben, amikor egy mondatot kétféle nemre is lefordít, „A fordítások megkülönböztetik a nemeket” felirat jelenik meg a fordítások fölött (2. Ábra[2]). A nemeket megkülönböztető fordítást további nemsemleges nyelvekre is tervezték kiterjeszteni (Kuczmarski 2018), ami jelen kutatás készülése közben magyar nyelvre is megtörtént. Így a Google Fordító jelenleg az „ő egy...” kezdetű mondatokat „he”-vel és „she”-vel kezdődő mondatokra is lefordítja angolul (2. Ábra). Mindazonáltal a Google Fordító magyar-angol fordításainak vizsgálata továbbra is indokolt lehet. Egyrészt a fordító azon funkciója, melynek segítségével teljes dokumentumokat lehet lefordítani, továbbra is egyféleképpen fordítja le a mondatokat, ahogy azt a fordító korábban tette. Másrészt magyarról az angolon kívüli más nemeket megkülönböztető (pl. francia, spanyol) nyelvre történő fordítás esetében még nem vezették be a nemeket megkülönböztető fordítást. A Google Fordító az angolt közvetítőnyelvként használja két nyelv között, így lehetséges, hogy a magyar-angol fordításokban előforduló nemi torzítások hatással vannak a magyar-francia, magyar-spanyol fordításokra és más nemeket megkülönböztető nyelvre történő fordításokra is (Prates és társai 2019). Harmadrészt azt a gépi tanuláson alapuló módszert, amit a Google Fordító hasznosít, számos más adatfeldolgozó program is használja,

[2] Forrás: Google Fordító [Képernyőkép] (https://translate.google.hu/?hl=hu&tab=uT#view=home&op=translate&sl=hu&tl=en&text=%C5%91%20egy%20orvos) (Utolsó megtekintés: 2020. április 23.)

köztük a Google Kereső, a Spotify vagy a Netflix is (Olson 2018). A legtöbb hasonló algoritmusnál nem lehet olyan egyszerűen kiküszöbölni a nemi vagy más jellegű torzításokat, ahogy azt a Google Fordító esetében meg lehetett tenni a két nemre való fordítás bevezetésével. Így a fordító esettanulmányon keresztül történő vizsgálata továbbra is releváns. A Google Fordító esete bepillantást enged abba, hogyan alakulnak ki társadalmi torzítások a gépi tanuláson alapuló algoritmusoknál, ezeket hogyan lehet vizsgálni és milyen lehetőségeink vannak a kijavításukra.

A Google Fordítót vizsgáló esettanulmányom arra a kutatási kérdésre próbál meg választ adni, hogy: Milyen mértékű a nemi alapú gépi torzítás a Google Fordítóban foglalkozások magyar-angol fordításánál? Következésképpen olyan mondatok angol fordítását vizsgáltam, mint „ő egy orvos” vagy „ő egy mérnök”. Ezen fordításokon keresztül térképeztem fel a nemi torzítás mértékét. A foglalkozások vizsgálatakor egy kiegészítő kutatást is végeztem, melyben a foglalkozásokhoz mellékneveket kapcsoltam. A kiegészítő kutatásban olyan mondatok fordítását vizsgáltam, mint „ő egy jó orvos”, „ő egy nagyon jó orvos”, „ő egy rossz orvos” és „ő egy nagyon rossz orvos”. Ezzel azt próbáltam meg feltérképezni, hogy a melléknevek hogyan változtatják meg az „ő” nemsemleges személyes névmás fordítását az egyes foglalkozásoknál.

Szakdolgozatomban először elméleti háttérként áttekintem a gépi tanuló algoritmusok működését, a bennük megjelenő társadalmi torzítások kialakulását, ezek vizsgálatának lehetőségeit és a torzítások kiküszöbölésének lehetséges módjait. Ezután áttérek a Google Fordító működésének ismertetésére és arra, hogy miért jelennek meg nemi torzítások egy fordítóprogramban. Majd ismertetem a Google Fordító fordításain végzett esettanulmányomat.

2. Elméleti áttekintés

2.1. Gépi tanulás és algoritmusok

Ahhoz, hogy áttekinthessük, hogy milyen társadalmi torzítások fordulhatnak elő a gépi tanuláson (angolul *machine learning*-en) alapuló algoritmusok kapcsán és miért, először is tisztáznunk kell, hogy mit értünk algoritmus és gépi tanulás alatt. Korábban ezeket a fogalmakat kizárólag az informatikában és társtudományaiban használták, de az utóbbi időben egyre gyakrabban fordulnak elő a mainstream médiában és a hétköznapi szóhasználatba is

bekerültek (Ságvári 2017; Burrell 2016). Ettől függetlenül továbbra is aluldefiniált kifejezések (Király 2020; Lovász 2018), amit egyfajta misztikum övez.

„A számítástechnika világában az algoritmus egy számítási folyamat absztrakt, formalizált [vagyis a számítógép, a programnyelv számára is olvasható] leírása" (Ságvári 2017: 64), amely valamilyen *input*ból (bemenetből) valamilyen *output*ot (kimenetet) hoz létre (Cormen és társai 2003; Király 2020; Lovász 2018). Tehát az algoritmus egy olyan precízen leírt eljárás, amely a bemeneti adatokból – amelyek lehetnek szövegek, számok, képek stb. – automatikusan előállítja a létrehozni kívánt eredményt.

Dolgozatom középpontjában a gépi tanuláson alapuló algoritmusok állnak. A gépi tanulást alkalmazó algoritmusok olyan algoritmusok, amelyeket akkor alkalmaznak, ha az adott feladatot túl nehéz lenne egy emberek által előre megírt programmal megoldani (Goodfellow és társai 2016). A hagyományos programozói feladatokat úgy oldották meg, hogy emberek a feladat megoldásához szükséges teljes folyamatot leprogramozták. A gépi tanuló algoritmusok ezzel szemben a beléjük táplált adatokból maguk próbálják meg felismerni, hogy a feladat elvégzéséhez milyen megoldásra lenne szükség (Google 2017). Úgy működnek, hogy egy emberek által létrehozott nagy mennyiségű adathalmazban tanulnak meg statisztikai valószínűségi alapon mintázatokat, összefüggéseket keresni (Barocas–Selbst 2016; Ságvári 2017). Például egy spam szűrő algoritmust olyan emaileket tartalmazó adathalmazon tanítanak, amelyeket emberek előre felcímkéztek aszerint, hogy az adott email „spam" vagy „nem spam". Az algoritmusnak ezen levelek alapján kell eldöntenie, hogy milyen tulajdonságokkal rendelkezik egy „spam" levél és egy „nem spam" levél tartalma, feladója és fejléce, és mi a kettő közötti különbség. A cél ezzel az, hogy az algoritmus a felismert különbségek alapján egy korábban fel nem címkézett emailről is el tudja dönteni, hogy az „spam" vagy „nem spam" (Burrell 2016).

A gépi tanuló algoritmusok működési mechanizmusukat tekintve három kategóriába sorolhatók: felügyelt, felügyeletlen és megerősítéses tanulást alkalmazó algoritmusok. A felügyelt tanulást alkalmazó algoritmusok esetében azt az adathalmaz, amelyen az algoritmust tanítják, emberek előre felcímkézik, majd ezután az adathalmazt két részre osztják. Azon részét, amelyen tanul a program, *training set*-nek (tanuló halmazna) hívják. Az algoritmus a training set-ben tanul meg összefüggéseket felismerni az adatok és az adatokhoz tartozó címkék között. Az algoritmus alapvető célja az, hogy a training set-ben felismert összefüggések ne csak a training set-re legyenek helytállóak és az algoritmus korábban nem látott adatokon

is jól működjön. Azt, hogy ezt milyen hatékonysággal teszi, az adathalmaz másik részével, a *test set*-tel (teszt halmazzal) tesztelik. Miután az algoritmus a training set-ben megtanulta felismerni a szükséges mintázatokat, leteszteltelik, hogy hogyan működik az általa korábban nem látott test set-en. A program készítői tudják, hogy milyen címkék tartoznak az egyes adatokhoz a test set-ben, de ezeket az algoritmusnak nem adják meg. Arra kíváncsiak ugyanis, hogy az algoritmus tudja-e reprodukálni a test set-ben az eredeti címkéket. Az algoritmus által kreált és az eredeti címkék összevetésével lehet mérni azt, hogy milyen hatékonyan működik az algoritmus (Goodfellow és társai 2016). Felügyelt tanulással lehet tanítani például az előbb ismertetett spam szűrő algoritmust. Az algoritmushoz használt adathalmaz minden emailjét emberek előre felcímkézik „spam" és „nem spam" címkékkel. Ezután adathalmazt két részre osztják. Az algoritmus a training set-ben lévő emailek és címkéjük alapján megtanulja felismerni a különbségeket a „spam" és „nem spam" levelek között. Ezen összefüggések alapján a test set-ben lévő minden egyes emailről eldönti, hogy azt „spam"-nek vagy „nem spam"-nek tartja. Ezt követően megvizsgálják, hogy a test set azon emailjeit, amit az algoritmus „spam"-nek tartott, eredetileg az emberek is „spam"-nek tartották-e. Ugyanezt az összevetést „nem spam" emailekkel is megteszik. Így mérik, hogy milyen hatékonyan működik a spam szűrő algoritmus.

A második tanulási módszert, a felügyeletlen tanulást alkalmazó algoritmusok esetében nem szükséges, hogy az adatok előre fel legyenek címkézve, mert az algoritmus az adatok struktúrájából vonja le következtetéseit. A felügyeletlen tanulás egyik alkalmazási területe a klaszterezés, amikor az algoritmus feladata az, hogy csoportokra bontsa az adatokat, de ezt nem emberek által előre gyártott címkék szerint teszi, hanem az adatok jellegzetességei alapján maga alkot csoportokat (Goodfellow és társai 2016). Ilyen például a topik modellezés, amikor az algoritmus különböző szövegeket csoportosít a témájuk alapján (Rangarajan Sridhar 2015).

A harmadik tanulási típusnál, a megerősítéses tanulásnál sincsenek előre felcímkézve az adatok, ebben az esetben a program a működésének értékeléséből, a visszacsatolásból tanul és fejleszti folyamatosan önmagát (Goodfellow és társai 2016). Erre a tanulási típusra jó példa az, amikor meg szeretnénk tanítani egy önvezető autót a parkolásra egy szimulált környezetben. A parkolást modellező szimulációban az autó kezdetben véletlenszerűen mozog, ha közel megy a parkolóhelyhez, megjutalmazzuk az algoritmust, ha nekimegy

valaminek, akkor megbüntetjük. Ezekből a megerősítésekből tanul meg végül parkolni a szimulációban (Arzt 2019).

Előbbiekben aszerint osztályoztuk a gépi tanuló algoritmusokat, hogy hogyan tanulnak összefüggéseket felismerni, de aszerint is kategorizálhatjuk őket, hogy ezt milyen célból teszik. A gépi tanulást sokféle feladat megoldására lehet alkalmazni, és ha megvizsgáljuk néhány alkalmazási módját, talán a gépi tanulás lényegét is jobban megérthetjük. Az egyik alkalmazási módja az úgynevezett klasszifikáció, amikor a programnak különböző kategóriákba kell rendeznie az egyes eseteket az adatokban található jellegzetességek alapján (Goodfellow és társai 2016). Erre jó példa a már említett spam szűrés, amikor a programnak „spam" és „nem spam" kategóriákba kell sorolnia a beérkező leveleket a levelek szövege, fejléce, küldője stb. alapján (Burrell 2016). De ilyenek azok az orvosi diagnózist támogató képfelismerő programok is, amelyek egy röntgenfelvétel vagy más képalkotó eljárással készült felvétel alapján szűrnek ki betegségeket (Burrell 2016; Ságvári 2017). A gépi tanulás egy másik alkalmazási területe a becslések és előrejelzések készítése (Goodfellow és társai 2016). Például hitelkérelmek elbírálása előtt annak eldöntése az ügyfelek élelmiszer-kiadásai alapján, hogy várhatóan visszafizetik-e a felvenni kívánt hitelt (Ságvári 2017). Vagy annak megbecslése a biztosítás megkötése előtt, hogy az adott ügyfél valószínűsíthetően mekkora kárigényt fog benyújtani és ez alapján mekkora biztosítási összeget kell befizetnie. A gépi tanulást alkalmazó algoritmusok ezenkívül használthatók hangfelvételek írott szöveggé alakítására vagy optikai karakterfelismerésre – arra az eljárásra, amikor az algoritmus egy fényképeken látható, kép formátumú feliratot szöveggé alakít. A gépi tanulásnak ezeken kívül még sok más alkalmazási területe van, végezetül a gépi fordítást emelném ki (Goodfellow és társai 2016). A szakdolgozatomhoz készült esettanulmányban a gépi fordításban megjelenő társadalmi torzításokkal foglalkozom.

Látható tehát, hogy a gépi tanulást rengeteg területen lehet alkalmazni, és a gépi tanuló algoritmusok az életünk nagy részét befolyásolják. Gépi tanuló algoritmusok határozzák meg, hogy internetes kereséseinkre milyen találatok és milyen információk jutnak el hozzánk vagy hogy az egyes online szolgáltatóknál mely hírek kerülnek a kiemelt hírek közé (Ságvári 2017). Az információfogyasztás és -szolgáltatás mellett az algoritmusok egyre nagyobb teret nyernek a befektetések, a hitelezés és a biztosítás, a bűnmegelőzés és az igazságszolgáltatás, a munkaerő-kiválasztás és -értékelés, valamint az ajánlási rendszerek és az online hirdetések területén is (Goodman–Flaxman 2016; Sandvig és társai 2014; Ságvári

2017). Egyre több területen bízzuk rá magunkat algoritmikus döntésekre és ezért fontos foglalkozni azzal a problémával, ha a gépi tanuló algoritmusok társadalmi igazságtalanságokhoz járulnak hozzá.

2.2. Algoritmikus diszkrimináció és algoritmikus torzítás

A gépi tanuló algoritmusok társadalmi igazságosságát vizsgáló kutatások általában az algoritmikus diszkrimináció (*algorithmic discrimination*) (Chen és társai 2018; Sandvig és társai 2014; Ságvári 2017) vagy az algoritmikus torzítás, gépi torzítás (*algorithmic bias*, *machine bias*) (Prates és társai 2019) kifejezéseket használják. Dolgozatomban az angol szakirodalmakban használt *bias* kifejezés megfelelőjeként a *torzítás* szót használom. Bár nagyobb hangsúlyt szentelek az algoritmusokban megjelenő torzításoknak és a saját kutatásomban is a gépi fordításban megjelenő társadalmi torzításokat vizsgálom, a témához szorosan kapcsolódik a diszkrimináció jelensége is, ezért az algoritmikus diszkriminációról is szó lesz. A következő két fejezetben áttekintem, hogy mit értünk diszkrimináció és torzítás alatt, ezek hogyan jelennek meg az algoritmusokban és hogyan kapcsolódnak össze.

2.2.1. Algoritmikus diszkrimináció

Ahhoz, hogy megértsük, mit jelent az algoritmikus diszkrimináció és hogyan jelenik meg, először tekintsük át, hogy mit értünk diszkrimináció alatt. Bár a diszkrimináció fogalma nagy múltra tekint vissza a szociológiában, nem csupán a szociológia sajátja. Ha diszkriminációról van szó, akkor beszélhetünk jogi értelemben vett diszkriminációról (Ságvári 2017) és a jogi szabályozás az algoritmusok kapcsán nem elhanyagolható, de szakdolgozatomban erre nem térek ki bővebben. Ehelyett a diszkrimináció szociológiai definíciójával foglalkozom. Szociológiai értelemben diszkrimináció alatt azt értjük, amikor bizonyos tulajdonságok (pl. bőrszín, nem, életkor, vallás, szexuális beállítottság stb.) alapján egyéneknek vagy emberek csoportjainak eltérő bánásmódban van részük és ezért eltérő lehetőségeik vannak függetlenül attól, hogy milyen egyéni képességekkel és kvalitásokkal rendelkeznek (Goodman–Flaxman 2016; Ságvári 2017). Ez az eltérő bánásmód nem feltétlenül jelent negatív bánásmódot, létezik pozitív diszkrimináció is, de amikor algoritmikus diszkriminációról van szó, akkor a cél a negatív diszkrimináció felismerése és kiküszöbölése. Algoritmikus diszkrimináció alatt azt értjük, amikor igazságtalanul vagy előítéletesen bánnak emberekkel egy automatikus döntési rendszer eredménye alapján (Mitchell 2019).

A diszkrimináció megvalósulhat közvetlen módon, amikor valakit vagy valakiket egy csoporthoz való tartozásuk miatt – többnyire szándékosan – eltérő módon kezelnek, ezt nevezik közvetlen diszkriminációnak. Az algoritmusok kapcsán ez akkor valósul meg, ha az algoritmus bemeneti adatait kifejezetten a csoportok közötti egyenlőtlenséghez kapcsolódó változók alkotják, például a nem vagy a kor (Chen és társai 2018). Ilyen például az idősekkel szembeni diszkrimináció Facebook-os álláshirdetések kapcsán. A The New York Times és a ProPublica közösen végzett kutatásuk (Angwin és társai 2017) során azt találták, hogy több cég is olyan álláshirdetést adott fel Facebook-on, amellyel csak a fiatalabb korosztályt targetálták, és ezzel megfosztották a megcélzott korosztályon kívüli idősebb személyeket attól a lehetőségtől, hogy tudomást szerezzenek az álláslehetőségről és jelentkezhessenek arra.

A diszkriminációnak nem csak ez a nyílt formája létezik, beszélhetünk közvetett diszkriminációról is. Ez esetben egy társadalmi csoporttal szemben már létező diszkrimináció további alkalmazásáról, intézményesüléséről van szó, olykor nem is szándékos módon (Chen és társai 2018; Ságvári 2017). Az algoritmusok kapcsán elsősorban ez utóbbi diszkriminációs folyamatról van szó, és nem arról, hogy az algoritmusokat eleve diszkriminatívnak szánják. Ha a társadalmi csoportok közötti különbségtétel nem célja az algoritmusnak, de mégis különböző módon bánik velük, az az algoritmus működésében rejlő torzítások miatt van. Az algoritmikus diszkrimináció oka ekkor az algoritmikus torzítás, ennek módjairól a következő fejezetben lesz szó.

2.2.2 Algoritmikus torzítás

Torzításról bármely olyan adatgyűjtési és -feldolgozási folyamat kapcsán beszélhetünk, ahol valamilyen szisztematikus hiba lép fel, legyen az tudományos kutatás vagy a gépi tanuló algoritmus által felhasznált adatok előállítása és feldolgozása. Azt, hogy egy adatgyűjtési és -feldolgozási folyamat során torzítás történt, mindig csak valamilyen viszonyítási alaphoz képest mondhatjuk. A torzítás ezen viszonyítási alaptól való eltérést jelenti (Freedman és társai 2005). Ez az eltérés sokféle okból keletkezhet. Probléma léphet fel az adatok gyűjtése, feldolgozása, felcímkézése során, de már az adatgyűjtés előtt is lehetnek torzítások az adatokban. Torzítások léphetnek fel amiatt, mert pontatlanul figyeljük meg a körülöttünk lévő világot, például azért, mert megfigyeléseinket előítéletek, sztereotípiák vezérlik, de ugyanúgy befolyásolja megfigyelőképességüknet az is, hogy a kiugró, atipikus eseteknek nagyobb figyelmet szentelünk; az, hogy a saját csoportunk tagjait sokfélének látjuk, ellentétben más

csoportok tagjaival, akiket hasonlónak; az, ha specifikus információkból túl általános következtetéseket vonunk le, túláltalánosítunk; vagy az, hogy azokat az információkat részesítjük előnyben, amelyek az előzetes meggyőződéseinket erősítik meg – ez utóbbit nevezzük megerősítéses torzításnak. Ezek, a szociálpszichológia által jól leírt torzítási fajták mind befolyásolhatják azon adatok minőségét, amelyen a gépi tanuló algoritmust tanítják. Ahogy az is, ha adataink nem reprezentálják megfelelően a valóságot, mert a mintánkba kerülő eseteket nem jól választottuk ki. Ha emberekről szóló adatokról van szó, akkor ilyen esetben előfordulhat, hogy egyes társadalmi csoportok felül, mások alul vannak reprezentálva a mintánkban és a programnak ilyen torz adatokon kell dolgoznia. Összefoglalva akkor léphet fel torzítás egy adatfeldolgozási folyamat során, ha valamilyen hiba csúszik az adatok kiválasztásába és feldolgozásába vagy már eleve valamilyen szempontból torzított adatokon dolgozunk. Ez a torzítás nem feltétlenül jár negatív következményekkel, például az optimizmus is egyfajta kognitív torzítás. A gépi torzítás alatt viszont szűkebb értelemben azt értjük, amikor egy programban, ami fontos társadalmi döntések alapjául szolgál, megjelenik a szisztematikus torzítás és az algoritmikus döntések következtében egyes társadalmi csoportok hátrányos helyzetbe kerülnek (Mitchell 2019).

Szakdolgozatom egyik kulcskérdése az, hogy az algoritmusok hogyan válhatnak torzítóvá. Hajlamosak vagyunk azt gondolni, hogy a számítógépek és programok hozta döntések objektívek, használatukkal kiiktatható az emberi szubjektivitás, előítélet, sztereotípia és hiba a döntéshozatalból (Barocas–Selbst 2016; Ságvári 2017). Azonban az algoritmusokat emberek írják, a működésükhöz szükséges adatokat emberek gyűjtik, rendszerezik, tisztítják és tárolják, a gépi tanuláson alapuló rendszereket emberek által létrehozott és emberekről szóló adatokon tanítják. Vagyis az emberi tényező nem iktatható ki az objektívnek hitt algoritmusok működéséből és így végül az általuk létrehozott eredményből sem (Mitchell 2019; Ságvári 2017). Azt, hogy miként torzíthat egy algoritmus, kétféle dolog okozhatja. Egyrészt lehetséges, hogy az algoritmus által feldolgozott adatokkal van a probléma és emiatt torzít (Ságvári 2017). Barocas és Selbst (2016: 671) ezt így fogalmazták meg: „egy algoritmus csak annyira lehet jó, amennyire az általa használt adatok azok” (saját fordítás). Másrészt előfordulhat, hogy az algoritmus működése vet fel problémákat (Ságvári 2017). Először azt az eshetőséget tekintem át, ha a gépi tanulást alkalmazó algoritmusok azért torzítanak, mert a bemenő adatokkal van probléma.

Egyrészt probléma lehet az algoritmus által feldolgozott adatokkal azért, mert megjelennek bennük a társadalomra jellemző egyenlőtlenségek, előítéletek, sztereotípiák (Goodman-Flaxman 2016). Például a training set-ben fellelhető esetek úgy vannak felcímkézve, hogy azok a korábban kialakult előítéleteket, sztereotípiákat tükrözik (Barocas–Selbst 2016). Ilyenkor, ha a program készítői nem tesznek lépéseket azért, hogy kiküszöböljék vagy mérsékeljék azokat, akkor a program megismétli ezeket az előítéleteket, sztereotípiákat, bizonyos esetekben fel is erősíti azokat, ami a diszkriminatív eljárások és a társadalmi igazságtalanság intézményesüléséhez vezet (Ságvári 2017). Így az objektívnek tartott algoritmusok részrehajló, emberek egy csoportját hátrányosan érintő döntéseket eredményeznek (Goodman-Flaxman 2016). Különösen aggasztó, ha a részrehajló bemeneti adatokon alapuló részrehajló, diszkriminatív kimeneti adatok maguk is bemeneti adatokká válnak, mert ilyenkor tovább erősödik a torzítás. Ez a folyamat a következőképpen zajlik le: az emberi torzításokat tartalmazó bemeneti adatokon tanuló algoritmus maga is torzításokkal teli kimeneti adatokat produkál, majd az emberek ezen kimeneti adatok szerint cselekednek, hoznak döntéseket, és ezen cselekedetek, döntések újabb bemeneti adatokká válnak a program számára. Mitchell (2019) ezt a folyamatot *feedback loop*-nak, visszacsatolási ciklusnak nevezi és példaként az előrejelzésen alapuló bűnmegelőzést hozza fel. Egy olyan bűnmegelőzési rendszert, amely előre jelzi, hogy várhatóan hol fordulhatnak elő bűncselekmények az alapján, hogy korábban mely területeken voltak letartóztatások. Nem az alapján, hogy hol *történtek* bűncselekmények, hanem hogy hol *jelentettek be* bűncselekményt. Hiszen a rendőrség csak azokról a bűncselekményekről tud, amit be is jelentettek, azokról nem, amit ugyan elkövettek, de nem jelentették és nem történt letartóztatás. Így a bűnmegelőző algoritmus azokra a területekre jelez várható bűncselekményeket, amelyeken már eleve többször fordulnak meg a rendőrök valamilyen emberi részrehajlás miatt. Ennek okán még nagyobb figyelem irányul ezekre a területekre, még több letartóztatás történik ott és az algoritmus még kritikusabbnak fogja ítélni azokat. Elindul egy visszacsatolási ciklus, ami a bemeneti adatokban rejlő részrehajlásból, előítéletekből, sztereotípiákból fakad.

Másrészt akkor is torzítás keletkezhet az adatokból, ha nem jól választottuk ki a bemeneti változónkat. Előfordulhat, hogy a bemeneti adatokban nincsen kifejezetten diszkrimináló változó, az algoritmus mégis hátrányban részesít egyes csoportokat. Ez azért lehetséges, mert a bemeneti adatokban van egy vagy több olyan „semleges” változó, ami

szorosan korrelál egy diszkrimináló változóval, és ezért a bemeneti adatokban helyettesíti azt, proxy változóként működik. Következtetni lehet belőle az egyén nemére, korára, faji hovatartozására és más diszkriminációhoz kapcsolódó jellemzőkre (Barocas–Selbst 2016; Chen és társai 2018). Például az irányítószám tipikusan egy ilyen proxy változó. Egyes amerikai pénzintézetek az irányítószám alapján megtagadhatják, hogy az adott ügyfél jelzálogkölcsönt vagy biztosítást kössön és ezzel impliciten hátrányos helyzetbe hozzák a fekete lakosságot és a lepusztult környékek lakóit (Chen és társai 2018). A pénzintézetek az irányítószámot a fizetőképesség indikátoraként használják annak érdekében, hogy hatékonyabban ki tudják szűrni a fizetésképtelen ügyfeleket. Ezzel az a probléma, hogy az irányítószám nem az egyén, hanem az adott irányítószám alatt lakó összes ember fizetőképességével függ össze, csak statisztikai alapon köthető az egyénhez. És így az irányítószám figyelembevételével olyan emberektől tagadhatják meg, hogy hitelt vegyenek fel vagy biztosítást kössenek, akik egyébként fizetőképesek lennének, és ezzel további hátrányban részesítenek eleve diszkriminált csoportokat.

Harmadrészt az algoritmus által használt adatokkal akkor is probléma van, ha azok nem reprezentálják megfelelően a valóságot. Súlyos következményei lehetnek annak, ha egyes társadalmi csoportok nincsenek megfelelően képviselve az adatokban, mert a gépi tanuláson alapuló algoritmusok ezekben a nem reprezentatív adatokban keresnek összefüggéseket, majd alkalmazzák a szélesebb társadalmi közegre. A probléma itt abból adódik, hogy olyan adatokból vonnak le következtetéseket a populációra vonatkozóan, amik nem írják le megfelelően a populációt. Ha túl nagy a különbség azon esetek között, amin a gépi tanuló algoritmust tanítjuk és azon esetek között, amire majd alkalmazni szeretnénk, akkor az algoritmus nem fog jól működni, torzítani fog, hiszen az általa felismert összefüggések csak egy nagyon speciális, a társadalmat nem jól reprezentáló mintában lesznek helytállóak (Ságvári 2017). Ilyenkor – informatikai kifejezéssel élve – a modell túltanulja magát, a training set-en jól fog működni, de a test set-en és a valós alkalmazása során nem (angolul ezt a jelenséget nevezik *overfitting*-nek) (Goodfellow és társai 2016). Ha az adatokban nincs megfelelően reprezentálva a társadalom, mert egyes társadalmi csoportok felül, mások alul vannak reprezentálva az adatokban, annak az a következménye, hogy az algoritmus egyes társadalmi csoportokat a többiekhez képest előnyben, másokat hátrányban fog részesíteni (Barocas–Selbst 2016).

A digitális adatokon nyugvó algoritmusok esetében problémás, hogy azok a csoportok, akik nem használnak digitális eszközöket és kevesebb digitális lábnyommal rendelkeznek, nem fognak megjelenni a training set-ben, és így a teljes társadalmat célzó algoritmusok esetében hátrányba kerülnek (Barocas–Selbst 2016). Ráadásul az esetükben már eleve egy sok szempontból hátrányos csoportról van szó, mivel a digitális egyenlőtlenség összefügg egy sor más egyenlőtlenséggel: gazdasági, kulturális és társadalmi egyenlőtlenségekkel (Nagy 2007). A természetes szövegeket feldolgozó algoritmusok esetében a kevesebb ember által beszélt nyelvek és dialektusok azok, amik kimaradnak a training set-ből. Például a legtöbb nyelvfelismerő rendszert nem tréningezik regionális dialektusokon, ennek eredményeképpen rendszeresen rosszul klasszifikálják a regionális dialektust használó szövegeket (Jurgens és társai 2017).

A három felsorolt ok közül akármelyik miatt is van probléma az adatokkal, a gépi tanuló algoritmus az adatokban lévő torzítások, hibák miatt torzítani fog. Ráadásul ezt a torzítást az algoritmus még fel is nagyíthatja azáltal, hogy a sztereotipizált tulajdonság (pl. életkor, nem) vagy annak proxy változója alapján hátrányban részesíti, lepontozza, a rendszerben hátrébb sorolja az egyént (Ságvári 2017). Erre jó példák azok az álláskereső portálok által alkalmazott algoritmusok, amelyek a felhasználók által feltöltött önéltrajzokat pontozzák és rangsorolják a munkaadók számára (Chen és társai 2018; Ságvári 2017). Chen és társai (2018) három olyan álláskereső portált vizsgált, ahol a munkaadók maguk kereshetnek álláskeresők önéletrajzai között különböző kulcsszavak és filterek segítségével. A vizsgálat során azt találták, hogy az algoritmusok enyhe mértékben hátrébb sorolták a női jelentkezőket a hasonló tapasztalattal rendelkező férfiakkal szemben. Hasonló kritika merült fel az Amazon saját használatra fejlesztett toborzó programja kapcsán (Dastin 2018; Schwarm 2018). A program előnyben részesítette a férfiakat a nőkkel szemben, mert nem az adott munkához kapcsolódó kifejezéseket preferálta, hanem az olyan kifejezéseket, amelyeket a férfiak használtak többször. Valamint lepontozta azokat az önéletrajzokat, amelyekben szerepelt az, hogy „női”, például úgy, hogy „a női sakk klub vezetője” (Dastin 2018). Ez azért történt, mert a modell olyan adatokon tanult, amely elsősorban férfiak önéletrajzából állt, vagyis a férfiak felül voltak reprezentálva a mintában (Schwarm 2018). Ezek az eredmények azért aggasztóak, mert amennyiben egy online toborzó program egy társadalmi csoport tagjait szisztematikusan hátrébb sorolja, a toborzók kisebb eséllyel olvassák el önéletrajzukat és így kisebb eséllyel

alkalmazzák őket. A társadalmi csoportokkal szemben torzító toborzó programokban a munkaerőpiaci diszkrimináció digitális változata jelenik meg (Chen és társai 2018).

Eddig arról volt szó, amikor a torzítás az algoritmus által feldolgozott adatokból származik, de probléma lehet magával az algoritmussal is. Az algoritmusokra a racionalitás, értéksemlegesség, objektivitás jellemző, ez eredményezi azt, hogy az algoritmikus rendszereket megbízhatónak tartjuk, megbízhatóbbnak, mint a részrehajlásra hajlamos emberi döntéseket. Azonban egy algoritmus megírása során számtalan olyan pont merülhet fel, ahol az algoritmus készítőinek értékalapú döntéseket kell hozniuk, és így ezek az értékítéletek a program működésébe is belekerülnek. Egy program végső soron igaz-hamis döntésekre, egyesekre és nullákra vezethető vissza, vagyis olyan elágazási pontokra, ahol a program készítőinek két ellentétes állítás közül kell választaniuk. Olyan ellentétpárokról van szó, mint jó-rossz, sok-kevés, nagyon-kicsit. Ilyenkor a programozó egyéni döntése az, hogy a bináris választási lehetőségekből melyiket választja. Az ilyen értékalapú döntéseknél sokszor nincs jó és rossz válasz, de a program írójának mégis mérlegelnie kell, hogy melyik döntés következményeit tartja kívánatosabbnak vagy kerülendőbbnek. Ez a mérlegelés elvezet az elsőfajú és másodfajú hiba kérdéséhez: a készítőknek el kell dönteniük, hogy az elsőfajú vagy a másodfajú hibát minimalizálják (Ságvári 2017). Például egy olyan orvosi diagnózist támogató algoritmus, amely röntgenfelvételek alapján ismer fel betegségeket, jobb, ha az az elsőfajú hibát preferálja, vagyis olyan esetben is betegnek diagnosztizálja a pácienst, ha az valójában nem beteg, hiszen így kiküszöbölhető, hogy ne részesüljenek megfelelő ellátásban azok, akik valóban betegek. Ezzel ellentétben a spam szűrés esetében a másodfajú hiba a megengedetteb, a szűrő inkább „nem spam”-ként kezel spam leveleket, minthogy fontos levelek a spambe kerüljenek (Mitchell 2019). Vannak azonban olyan esetek, amikor nehezebb eldönteni, hogy egy algoritmikus döntésnél mi a jó: az, ha túl szigorú és az elsőfajú hibát preferálja, vagy az, ha túl megengedő és a másodfajú hibát részesíti előnyben (Ságvári 2017).

A gépi tanuló algoritmusok esetében további nehézséget okoz, hogy bár kezdetben a programot emberek írják, a gépi tanulási folyamat automatikusan történik, nem lehet előre látni az összes lehetséges kimenetet, a tervezők nem látnak bele a folyamat lefolyásának menetébe, az algoritmus úgy működik, mint egy fekete doboz (Ságvári 2017). Ez megnehezíti a programozói döntések meghozatala előtti mérlegelést és az egyenlőtlenségek kiküszöbölését is, ugyanis nem lehet előre tudni, hogy mely bemeneti adatok esetében lesz diszkriminatív az output (Barocas–Selbst 2016). Különösen igaz ez a felhasználói inputokra

támaszkodó algoritmusokra, mint például a célzott hirdetésekre, ahol minden egyes felhasználó inputok egyedi kombinációját jelenti, és a program alkalmazása és hatásának megvizsgálása előtt nem lehet megmondani, hogy diszkriminációt okoz-e. Ahol sem a felhasználók, sem a program készítői nem látnak bele az automatikus döntések mögötti mélyebb folyamatokba és ezek a folyamatok az emberek számára értelmezhetetlenek, ott ember és gép felelőssége közötti határvonal elmosódik (Ságvári 2017).

Összességében elmondható, hogy a gépi tanuló algoritmusok esetében a torzítás három különböző szinten születhet meg. Az első két szinten a torzítás forrása a training set-et alkotó adatok keletkezése és kiválasztása. A rendelkezésünkre álló adatbázisok, amiket a gépi tanuláson alapuló modellek használnak, szükségszerűen leegyszerűsítve írják le a körülöttünk lévő világot (Barocas–Selbst 2016). Ez okozza azt, hogy ha adatainkat nem körültekintően választjuk ki, ezekben az adatbázisokban torzítások keletkezhetnek és előítéletes, diszkriminatív rendszerek alapjául szolgálhatnak. Azonban előítéletek, sztereotípiák anélkül is lehetnek adatainkban, ha a lehető legkörültekintőbben választottuk ki őket, egyszerűen azért, mert az előítéletek és a sztereotípiák részei a minket leíró valóságnak. A torzítások keletkezésének első szintjét az jelenti, amikor a társadalomban korábban létrejött előítéletek, sztereotípiák bekerülnek a mintánkba akár egy diszkrimináló változó akár egy proxy változó használatán keresztül. A második szinten, az adatok gyűjtése során újabb torzítások léphetnek érvénybe: a pontatlan megfigyelés, a túláltalánosítás, a megerősítéses torzítás, a kiválasztási torzítás és más, korábban tárgyalt torzítási fajták. A harmadik szintnek azt tekintem, amikor az algoritmus a működési mechanizmusából vagy a programozói döntésekből adódóan felerősíti ezeket a torzításokat. Majd ezek a torzítások az algoritmusokon alapuló döntések következtében intézményesülnek és akár újabb bemeneti adatokká válhatnak. Ezáltal a folyamat ciklikussá válhat, kialakulhat egy visszacsatolási ciklus. Azt, hogy a három szint közül pontosan melyiken vagy melyikeken jelenik meg torzítás és így pontosan mi okozza azt, hogy egy algoritmikus rendszer igazságtalanul, előítéletesen, diszkriminatívan bánik emberekkel, nehéz megmondani, különösen akkor, ha nincs hozzáférésünk az adatokhoz vagy az algoritmus kódjához (Sandvig és társai 2014; Ságvári 2017). Ennek ellenére az algoritmikus torzítás és az algoritmikus diszkrimináció meglétét lehet vizsgálni. Azt, hogy a társadalmi igazságosság szempontjából az algoritmusok átvizsgálásának milyen lehetőségei vannak, a következő fejezetben tekintem át.

2.3. Az algoritmikus torzítás és diszkrimináció mérése és kiküszöbölése

Annak ellenőrzésére, hogy egy algoritmus igazságosan működik-e, a klasszikus diszkrimináció kutatások audit módszerére épülő algoritmusaudit módszerét szokták alkalmazni (Chen és társai 2018; Sandvig és társai 2014; Ságvári 2017). Sandvig és társai (2014) az algoritmusaudit öt típusát szedték össze: (1) kódaudit, (2) noninvazív felhasználói audit, (3) *scraping* audit, (4) *sock puppet* audit és (5) a felhasználóknak kiszervezett (angolul *crowdsourced*) vagy kollaboratív audit.

A kódaudit módszer lényege, hogy a kutatók közvetlenül az algoritmus kódját vizsgálva ellenőrzik, hogy az diszkriminatívan működik-e. Ennek alkalmazása a gyakorlatban több okból is nehézkes. Egyrészt a legtöbb algoritmus kódjához nem lehet hozzáférni, mert nem publikus. Az algoritmusok szellemi tulajdonnak és üzleti titoknak minősülnek, ha nyilvánosan elérhető lenne a kódjuk, az hátrányba hozná az algoritmust használó céget a versenytársakkal szemben. Az algoritmusok kódját nem csak üzleti, de biztonsági okokból sem hozzák nyilvánosságra. Ha nyilvánosan elérhető lenne a kód, akkor könnyen „meghekkelhető" lenne a program. Bár vannak nyílt forráskódú programok, a teljes kód az előző ok miatt esetükben sem nyilvános. Másrészt a kódaudit módszerben problémát jelent az is, hogy az, hogy egy algoritmus diszkriminatív-e, nem csak a kódja miatt lehet, hanem a bemenő adatok miatt is. Sok algoritmus felhasználói inputtal működik, a felhasználói input hatását pedig csupán a kód ellenőrzésével nem lehet vizsgálni. Az algoritmikus torzítás és diszkrimináció vizsgálata tehát az algoritmusok perszonalizált, személyre szabott működése miatt sem egyszerű (Sandvig és társai 2014; Ságvári 2017).

A második auditálási módszer, a noninvazív felhasználói audit során a kutatók a felhasználóktól kérdőív segítségével szereznek információkat arról, hogy milyen tapasztalataik voltak az adott algoritmussal kapcsolatban, tapasztaltak-e diszkriminációt. Ebben a módszerben a mintavétel és a megbízhatóság jelenthet problémát: kérdéses, hogy mennyire lehet egy algoritmus működéséről pontos információkat szerezni felhasználók szubjektív beszámolói alapján (Sandvig és társai 2014; Ságvári 2017). Továbbá nem biztos, hogy mindig a felhasználók tudomására jut, ha egy algoritmus által diszkrimináció áldozataivá válnak (Dastin 2018).

A harmadik auditálási módszer a scraping audit, amikor a kutatók automatizált módon gyűjtenek adatot egy algoritmus működéséről. Ez a módszer akkor alkalmazható, ha az adatok

szabadon elérhetőek az interneten, de bizonyos esetekben jogi akadályokba ütközhet (Sandvig és társai 2014; Ságvári 2017).

A negyedik módszer, a sock puppet audit lényege, hogy a kutatók a diszkrimináció meglétét magukat felhasználóknak kiadva vizsgálják. Például felhasználói profilokat hoznak létre. Ennek a módszernek az egyik hátránya, hogy a kutatók beleavatkoznak a program működésébe, hiszen hamis adatokat töltenek fel egy olyan rendszerbe, aminek a működése adatokon alapul. A sock puppet audit esetében etikai és jogi akadályok is felmerülhetnek. (Sandvig és társai 2014; Ságvári 2017).

Az ötödik módszer hasonló a sock puppet audithoz, azzal a különbséggel, hogy a programmal nem a kutatók, hanem valódi felhasználók lépnek interakcióba. A kutatóknak ehhez önkénteseket vagy fizetett tesztelőket kell keresniük, ami költséges és nagy koordinációt igénylő feladat (Sandvig és társai 2014; Ságvári 2017).

Az algoritmusok igazságosságának vizsgálata tehát gyakran módszertanilag, jogilag és etikailag sem egyszerű feladat. További nehézséget jelent, hogy az algoritmusaudit során nehezen deríthető ki, hogy pontosan mi okozza az igazságtalanságot (Barocas–Selbst 2016; Chen 2018), és így az algoritmus kijavításának lehetősége sem egyszerű. A fekete dobozként működő algoritmusok esetében a programozók sem feltétlenül tudják előre, hogy egyes változtatásokkal kevésbé diszkriminatív vagy diszkriminatívabb lesz-e a program, erre az algoritmus utólagos auditálása adhat választ (Barocas–Selbst 2016). Mitchell (2019) az algoritmikus torzítás enyhítésének két lehetőségét említi meg. Az egyik módszert *debias*-nak vagy *unbias*-nak szokták hívni, ebben az esetben megpróbálják eltávolítani a modellből azt, ami a torzítást okozza. A másik lehetőség az, ha az algoritmust úgy tervezik meg, hogy az kezelni tudja, ha valamilyen előre meghatározott csoporttal szemben torzítás, diszkrimináció lépne elő. A Google Fordító nemeket megkülönböztető fordítása, vagyis amikor mindkét nemre lefordítja az algoritmus a nemsemleges személyes névmásokat, ez utóbbi típusú javítása az algoritmusnak.

Ahhoz, hogy ki lehessen javítani egy algoritmust, tisztában kell lenni azzal, hogy hogyan kellene működnie egy igazságos algoritmusnak, milyen változók mentén megengedhető és milyen változók mentén nem megengedhető, hogy eltérően kezeljen embereket. Például egy hitelképességet bíráló algoritmus esetében jogosnak tartjuk, ha a vagyoni helyzet alapján dönt, de jogtalannak, ha az ügyfél irányítószáma alapján. Az, hogy mely változók mentén nem szabad, hogy diszkriminatív legyen az algoritmus, a kutatók és a készítők döntésén múlik.

Például a szövegfeldolgozó algoritmusok esetében a nemi torzítás kiküszöbölése mellett szóba jöhet a vallási, faji, dialektikai torzítás kiküszöbölése. Ságvári (2017: 68) kiemeli, hogy „[a]z ilyen jellegű korrekció ugyanakkor pozitív diszkriminációnak tekintendő, amely újabb módszertani, jogi és etikai kérdések sorát veti fel".

2.4. Google Fordító

Ahogy az előbbi fejezetben ismertettem, az algoritmikus torzítás és diszkrimináció mérése módszertani szempontból a legtöbbször nem egyszerű feladat. A Google Fordító vizsgálatával azonban betekintést lehet nyerni a gépi tanulásban előforduló társadalmi torzításokba, ugyanis a legtöbb gépi tanuló algoritmussal ellentétben a Google Fordító működése nem függ a felhasználói inputtól – vagyis bármely két felhasználó számára ugyanazt a mondatot ugyanúgy fordítja le –, és ezáltal a Google Fordító vizsgálatát az átlag gépi tanuló algoritmusokhoz képest egyszerűbb megvalósítani. Mielőtt rátérnék arra, hogy a társadalmi igazságosság szempontjából hogyan lehet vizsgálni a Google Fordítót és milyen kutatásokat végeztek eddig a témával kapcsolatban, tekintsük át, hogy a Google Fordító, mint gépi tanuláson alapuló rendszer, milyen működési elven alapszik és miért jelenhetnek meg benne társadalmi torzítások.

Az automatizált gépi fordításban évtizedek óta a statisztikai gépi fordítás (*Statistical Machine Translation*) az uralkodó paradigma (Laki 2018; Prates és társai 2019; Wu és társai 2016), a Google Fordító is ezt a módszert követi (Wu és társai 2016). A statisztikai gépi fordításhoz, ahogy más gépi tanuláson alapuló rendszerekhez is, rengeteg adatra van szükség. A gépi fordítóprogramok esetében az adatot ebben az esetben olyan szövegkorpuszok alkotják, amelyekben az egyes szövegek legalább két nyelven elérhetőek, így az adatokban minden egyes mondatnak megtalálható annak fordítása is (Laki 2018). A program ezen mondatpárok összehasonlításából felügyelt módon tanulja meg felismerni a fordításokhoz szükséges összefüggéseket (Cho és társai 2019; Laki 2018). A Google Fordító az interneten elérhető több százmillió, több nyelvre lefordított szövegből tanul (Kuczmarski 2018). Ez kezdetben elsősorban az ENSZ és az Európai Unió által publikált anyagokat jelentette, amelyek több nyelven is elérhetőek voltak (Prates és társai 2019), de 2014-től kezdve a felhasználók is hozzájárulhatnak a fordítások pontosításához a Google Fordító Közösség kezdeményezésen keresztül: értékelhetik a fordításokat és maguk is generálhatnak új fordításokat (Prates és társai 2019; Kelman 2014).

Napjainkban a statisztikai gépi fordítás egyik legeredményesebb modellje a mesterséges neurális hálózatokon[3] alapuló szóbeágyazási modell (Laki 2018), 2016 óta a Google Fordító is ezt alkalmazza (Wu és társai 2016). A szóbeágyazási modellt használó programok úgy működnek, hogy figyelik a szövegkorpuszban található szavak együttjárását, azt, hogy az egyes szavak milyen más szavak társaságában fordulnak elő a leggyakrabban. Majd ez alapján meghatározzák, hogy egyes szavak jelentése közel vagy távol áll egymástól (Bolukbasi és társai 2016; Laki 2018; Olson 2018). A szavakat egy vektortérben ábrázolják, ahol azok a szavak, amelyek hasonló jelentéssel bírnak, közel helyezkednek el egymáshoz, míg azok a szavak, amelyek jelentése lényegesen eltér egymástól, távolabb helyezkednek el egymástól (Bolukbasi és társai 2016; Laki 2018). Mint minden gépi tanuláson alapuló modell, a szóbeágyazási modell működése is nagyban függ a bemenő adatok minőségétől és amennyiben az adatokban társadalmi torzítások vannak, úgy a modell át fogja venni azokat (Olson 2018).

Az utóbbi időben többen (Bolukbasi és társai 2016; Olson 2018) felfigyeltek arra, hogy a szóbeágyazási modellek hajlamosak a nemi torzítások átvételére, ugyanis megtanulják azokat a nemi sztereotípiákat, amik az adatokban (és a társadalomban) vannak. Egy jól működő szóbeágyazási modelltől azt várjuk, hogy összekösse a „nő” és az „anya”, valamint a „férfi” és az „apa” szavak jelentéseit. Azonban az már kétségeket vet fel, ha az olyan szavakat, amelyek önmagukban nem utalnak sem nőre sem férfira, mint az „orvos” vagy az „erős”, szorosan összeköti a „férfi” szóval, vagy a „recepciós”-t és a „szép”-et a „nő” szóval. Pedig, ha abban a szövegkorpuszban, amin a modellt tanítják az „orvos” szó környezetében gyakran fordul elő a „férfi” szó, a „nő” pedig ritkán, akkor a szóbeágyazási modell az „orvos”-t a „férfi”-vel fogja összekötni (Olson 2018). A szóbeágyazási modellt használó Google Fordító ezért ítélte hímneműnek az „orvos”-t és fordította úgy azt a mondatot, hogy „ő egy orvos” úgy, hogy „he is a doctor”, mielőtt a magyar nyelvre is bevezették volna a kétnemű fordítások lehetőségét. Az olyan sztereotipikus elképzelések, amik az orvosokat és az erőt a férfiakkal, a recepciósokat és a szépséget a nőkkel azonosítják, bármely olyan algoritmikus döntés kapcsán visszaköszönhetnek, ahol szóbeágyazási modellt vagy más természetes szövegeket feldolgozó modellt alkalmaznak. Ezek a modellek a bemeneti adatokból megtanulják a nemekhez – akár

[3] A mesterséges neurális hálózat egy mélytanulásos módszer, a mélytanulás (*deep learning*) a gépi tanulás egy speciális fajtája (Goodfellow és társai 2016).

impliciten – kötődő asszociációkat. Ráadásul a szóbeágyazási modell ezeket a torzításokat fel is nagyíthatja (Bolukbasi és társai 2016).

Kutatások (Prates és társai 2019; Schiebinger 2014) azt is megfigyelték, hogy a Google Fordító hajlamos arra, hogy többször fordítson (például foglalkozások esetében) hímnemre, mint nőnemre. Ennek az az oka, hogy azokban az internetről szedett korpuszokban, amin a programot tanítják, sokkal többször fordulnak elő férfiakra utaló szavak, mint nőkre. Többször szerepelnek bennük hímnemű személyes névmások (pl. „he"), mint nőnemű személyes névmások (pl. „she") (Bolukbasi és társai 2016; Schiebinger 2014). Így a fordításokban előforduló különbségekben szerepet játszik az is, hogy a training data-ban a nők alul vannak reprezentálva.

Az előbbiekben áttekintettem, hogy miből származhatnak nemi torzítások a Google Fordítóban és általánosságban a gépi fordításban. Ezek jól lefedik az algoritmikus torzítás korábban ismertetett eseteit, szintjeit. Egyrészt azok a szövegkorpuszok, amelyekben a program a fordításokhoz szükséges összefüggéseket keresi, társadalmi előítéleteket, sztereotípiákat tartalmaznak. Másrészt ezek a szövegek nem reprezentálják megfelelően a lakosságot, ugyanis az internetes tartalmakban a férfiak felül vannak reprezentálva. Harmadrészt a Google Fordító korábbi verziója a nemsemleges nyelvekről (pl. magyarról) a nemeket megkülönböztető nyelvekre (pl. angolra) történő fordítás esetében egy bináris választási lehetőség előtt állt: vagy hímneműre (pl. „he"-re) vagy nőneműre (pl. „she"-re) kellett fordítania egy alapvetően nemsemleges kifejezést (pl. „ő egy orvos").

Eddig tudomásom szerint két tanulmány (Cho és társai 2019; Prates és társai 2019) született a Google Fordítóban megjelenő, foglalkozásokhoz kapcsolódó nemi torzításról. Jelen esettanulmány mintájául – bizonyos pontokon felülvizsgálva azt – Prates és társai (2019) esettanulmánya szolgált. A tanulmány keletkezésének idejében a Google Fordító még egy nemsemleges nyelv esetében sem ajánlotta fel a nemeket megkülönböztető fordítást. A szerzők 1019 foglalkozás fordítását vizsgálták 12 nemsemleges nyelv angol fordításainál – köztük magyar-angol fordításoknál. Olyan mondatokat fordíttattak le a Google Fordítóval, mint „ő egy orvos", ahol az „ő", egy nemsemleges személyes névmás, az angol fordításban lehetett a nemre utaló „he" vagy „she" is[4]. Mivel 12 nyelv fordításait vizsgálták, az angol fordítások nemi torzítását összességében tudták vizsgálni. A foglalkozások vizsgálata mellett

[4] Némely nyelvek esetén a fordítások között előfordult a semleges „it" névmás, de magyar-angol fordításoknál ez nem fordult elő.

egy kiegészítő kutatást is végeztek, amelyben 21 melléknév fordításait elemezték (pl. „ő boldog”, „ő szomorú”). Azt találták, hogy a nemi torzítás a foglalkozások mellett melléknevekre is kiterjed. Cho és társai (2019) tanulmánya a Google Fordítón kívüli két másik fordítóprogram – a Naver Papago és a Kakao fordító – vizsgálatával egészítette ki a gépi tanuláson alapuló fordító algoritmusokban megjelenő nemi torzítás vizsgálatát. A három fordítóprogram működését elemezték és vetették össze foglalkozások és melléknevek koreai-angol fordításán keresztül. Mivel az ismertetett két tanulmány az első fordítóprogramokkal társadalomtudományi szempontból foglalkozó kutatások között van, módszertanuk kiindulópontot jelenthet más további, hasonló témával foglalkozó kutatás számára.

3. Módszertan

Ebben a fejezetben a Google Fordítóról készített esettanulmány módszertanát fejtem ki. Esettanulmányomban azt vizsgálom, hogy milyen mértékű a foglalkozásokhoz kapcsolódó nemi torzítás a Google Fordító programban, olyan mondatok angolra történő fordításánál, mint „ő egy orvos” vagy „ő egy mérnök”. Arra a kutatási kérdésre keresem a választ, hogy milyen mértékű a nemi alapú gépi torzítás a Google Fordító programban foglalkozások magyar-angol fordításánál. Ehhez először definiálom, hogy mit tekintek nemi torzításnak a Google Fordítónál. Majd ismertetem a vizsgált foglalkozások kiválasztásának módját és azok szisztematikus fordításának elkészítését. Ezután kifejtem, hogy az elkészült fordítások alapján milyen módszerrel mértem a nemi torzítás mértékét. Majd az esettanulmány módszertanáról szóló fejezet végén ismertetem a foglalkozásokhoz kapcsolt melléknevekről szóló kiegészítő kutatásomat, amelyben azt vizsgáltam, hogy a „jó”, „nagyon jó”, „rossz”, „nagyon rossz” jelzők hogyan befolyásolják a foglalkozások fordításának nemét.

3.1. Nemi torzítás

Ahhoz, hogy mérni tudjuk a nemi torzítás mértékét, először is definiálnunk kell, hogy mit jelent a nemi torzítás a Google Fordítónál. Torzításról általánosságban akkor beszélünk, ha szisztematikus eltérést tapasztalunk valamilyen létező vagy elméleti alapfelvetéstől (Freedman és társai 2005). Egyesek szerint a fordítóprogramoknál elméleti alapnak azt tekinthetjük, ha a fordítóprogram egyenlő arányban fordítja a foglalkozásokat hímnemre és nőnemre, vagyis az a kívánatos cél, hogy a hímnemű és nőnemű fordítások aránya 50-50%

legyen[5] (Cho és társai 2019, Prates és társai 2019). Azonban ettől az 50-50%-os aránytól való eltérés még nem feltétlenül jelenti azt, hogy torzít az algoritmus. Amennyiben azt tekintjük torzításnak, ha a fordító eltér a hímnemű és nőnemű személyes névmások 50-50%-os arányától, akkor nem vesszük figyelembe például azt, hogy az egyes foglalkozásokat nem egyenlő arányban végzik nők és férfiak. Ezért tanulmányomban nem azt tekintem torzításnak, ha nem egyenlő arányú a hímnemre és nőnemre fordított foglalkozások aránya és nem azt tekintem ideális fordítónak, ami 50-50%-ban fordít hímnemre és nőnemre. Ehelyett ideális fordítónak azt tartom, ami a valós férfi-nő arányt tükrözi vissza és az ettől való eltérést tekintem torzításnak. Hasonló megoldást követtek Prates és társai (2019) is a Google Fordítót vizsgáló tanulmányukban.

Ahhoz, hogy egy fordítóprogram a foglalkozásokhoz kapcsolódó valós férfi-nő arányt tükrözze vissza, a training set-nek, amin a program tanul, szintén a valós arányt kell visszatükröznie. Azonban annak meghatározása, hogy mit tekintünk valós férfi-nő aránynak foglalkozások esetében, korántsem egyértelmű. Tekinthetjük valós alapnak az egyes foglalkozásokat végző férfiak és nők arányát, de azt is, hogy a társadalom hogyan vélekedik az egyes foglalkozásokról, melyeket tartja férfiasnak és melyeket nőiesnek. Éppen ezért tanulmányomban a fordításokat kétféle mutatóval hasonlítom össze, amik a foglalkozásokhoz kapcsolódó „valós” férfi-nő arányt hivatottak mérni: a foglalkozásokat végző férfiak és nők arányával a társadalomban és azzal, hogy a társadalom mely foglalkozásokat tartja férfiasnak és nőiesnek.

3.1.1. A foglalkoztatottak férfi-nő aránya

Tanulmányukban Prates és társai (2019) a foglalkozások fordításait az egyes foglalkozásokat végző nők és férfiak arányával vetették össze. Mivel 12 nyelv angol fordításait vizsgálták, a foglalkozások fordításának „he”-„she” arányát az amerikai Munkaügyi Statisztikai Hivatal (Bureau of Labour Statistics) foglalkozásstatisztikai adataival hasonlították össze. Ugyanakkor kétnyelvű fordítások esetében nemcsak a célnyelvi társadalomban – jelen esetben az amerikai társadalomban – lévő férfi-nő foglalkozási aránnyal való összehasonlítás lehet indokolt, hanem a forrásnyelvi társadalommal való összevetés is. Hiszen az a torzítás, ami a fordítóban keletkezik, a training set-ben lévő forrásnyelvi és célnyelvi szövegek miatt is létrejöhet.

[5] Azzal, hogy a Google Fordító bevezette egyes nemsemleges nyelvek – köztük a török és a magyar – esetében, hogy a fordítóprogram mindkét nemre lefordítsa az eredetileg nemsemleges mondatokat, tulajdonképpen ezt az 50-50%-ot érte el.

Magyar-angol fordításoknál a torzítás keletkezhet a training set-ben lévő magyar és angol szövegek miatt is. Ugyanis az eredeti szövegek íródhattak magyarul és angolul is, és így tartalmazhatják a magyar társadalomra jellemző nemi eltéréseket és sztereotípiákat és az angol szövegekben meglévő nemi eltéréseket és sztereotípiákat is. A Google Fordító tanítására eredetileg az ENSZ és az Európai Unió által publikált szövegeket használtak (Prates és társai 2019), így a szövegek nagyrésze eredetileg valószínűsíthetően angol volt, ami az amerikai foglalkozási statisztikákkal való összevetést indokolná. Azonban 2014-től kezdve felhasználóktól szerzett adatokra is támaszkodnak (Prates és társai 2019; Kelman 2014), ami a forrásnyelvi, jelen esetben a magyar adatokkal való összevetést is indokolttá teszi. Tanulmányomban ezért a foglalkozások fordítását mind magyar, mind amerikai foglalkozásstatisztikai adatokkal összevetem.

Az egyes foglalkozások magyarországi férfi-nő arányával való összevetéshez a KSH 2011-es népszámlálási adatait használtam (Népszámlálás 2011). A népszámlálási adatokban a foglalkozások a KSH által kidolgozott FEOR'08 kategóriákba voltak sorolva. A FEOR rendszerben a foglalkozások kategóriákba és ahhoz tartozó kódokba vannak sorolva (Központi Statisztikai Hivatal 2010). A 2011-es népszámlálás alapján mind a 485 FEOR kategóriában dolgozó férfiak és nők számáról volt adatom. A nemi torzítás méréséhez ezzel a férfi-nő aránnyal hasonlítottam össze azt, hogy az egyes foglalkozásokat a Google Fordító hímnemű vagy nőnemű személyes névmással fordította-e.

A fordításokat a magyar foglalkozásstatisztikai adatok mellett amerikai foglalkozásstatisztikai adatokkal is összehasonlítottam. Azért, hogy össze tudjam hasonlítani a fordítások amerikai adatoktól való eltérését a fordítások 2011-es magyar adatoktól való eltérésével, az összehasonlításhoz 2011-es amerikai adatokat vizsgáltam. Az elemzésembe a Bureau of Labor Statistics (BLS) 2011-es amerikai kérdőíves népességfelmérésen (Current Population Survey) alapuló adatait vontam be, amelyben az egyes foglalkozási kategóriákban foglalkoztatott száma és a foglalkoztatott nők aránya szerepel (Bureau of Labor Statistics 2011). A közölt adatokkal kapcsolatban felmerült három probléma. Egyrészt a BLS nem közölte a nők arányát azoknál a foglalkozási kategóriáknál, ahol a foglalkoztatottak száma kevesebb 50 ezernél, így ezeket a foglalkozási kategóriákat nem tudtam bevonni az elemzésembe, ez 576-ból 205 foglalkozási kategóriát jelentett. Így a BLS kimutatásaiból 371 foglalkozási kategória férfi-nő arányát tudtam figyelembe venni. Másrészt a BLS a katonai foglalkozásokra vonatkozóan egyáltalán nem közölt adatokat, így a katonai foglalkozások fordításait csak a

magyarországi férfi-nő aránnyal tudtam összehasonlítani. Harmadrészt a BLS adataiban a foglalkozások más struktúrát követnek, mint a magyar FEOR kategóriák. A BLS a SOC (Standard Occupational Classification) rendszert használja, ami kategóriákba rendezi az egyes foglalkozásokat. Ahhoz, hogy az elemzés során össze tudjam hasonlítani a magyar és az amerikai férfi-nő arányhoz mért torzítást, meg kellett feleltetnem a FEOR és a SOC rendszert egymásnak. Ebben a nemzetközi szinten használt foglalkozásokat kategorizáló ISCO (International Standard Classification of Occupations) rendszer nyújtott segítséget (Bureau of Labor Statistics 2015; Központi Statisztikai Hivatal é.n.[a]). Az ISCO kódokon keresztül minden FEOR kategóriához hozzá tudtam rendelni a hozzátartozó hivatalos SOC kategóriákat, erre látható néhány példa az 1. Táblázatban. Ez nem csak azt könnyítette meg, hogy össze tudjam hasonlítani a fordító magyar és amerikai adatokhoz mért torzítását, hanem a fordítandó foglalkozások kiválasztásánál is nagy segítséget nyújtott, erről részletesebben *3.2. A foglalkozások kiválasztása* című fejezetben lesz szó.

FEOR kód	FEOR elnevezés	ISCO kód	ISCO elnevezés (magyarul)	ISCO elnevezés (angolul)	SOC kód	SOC elnevezés
2117	Vegyészmérnök	2145	Vegyészmérnökök	Chemical engineers	17-2041	Chemical Engineers
2624	Elemző közgazdász	2631	Közgazdászok	Economists	19-3011	Economists
2726	Színész, bábművész	2655	Színművészek	Actors	27-2011	Actors
3335	Látszerész	3254	Látszerészek	Dispensing opticians	29-2081	Opticians, Dispensing

1. Táblázat: Néhány példa arra, hogy a FEOR kategóriáknak hivatalosan milyen SOC kategória felel meg.

3.1.2. A foglalkozásokkal kapcsolatos attitűdök

A különböző foglalkozások fordításait nem csak magyar és amerikai foglalkozásstatisztikai adatokkal hasonlítottam össze, hanem azzal is, hogy az egyes foglalkozásokat mennyire tartja a társadalom nőiesnek vagy férfiasnak. Ugyanis a foglalkoztatottak férfi-nő arányán kívül ez a foglalkozásokkal kapcsolatos attitűd is befolyásolhatja azt, hogy a fordítóprogramok tanításához használt szövegekben mely foglalkozásoknál és milyen mértékben jelennek meg a nemi sztereotípiák és eltérések. Az egyik legtöbbet hivatkozott foglalkozásokkal és nemekkel kapcsolatos attitűdöt mérő kutatás Shinar 1975-ös tanulmánya, melyben egy 8 pontos skálán mérte, hogy mennyire tartják a megkérdezettek férfiasnak, semlegesnek vagy nőiesnek az

egyes foglalkozásokat. Bár több Shinar tanulmányára épülő kutatás (Beggs–Doolittle 1993; Couch–Sigler 2001) is készült azóta, tudomásom szerint sem az USA-ban, sem Magyarországon nem készült olyan felmérés a közelmúltban, ami reprezentatív mintán mérte volna a foglalkozásokkal és nemekkel kapcsolatos attitűdöt. Ezért a fordításokat az amerikaiak foglalkozással kapcsolatos attitűdjével nem tudtam összehasonlítani. Viszont a magyar lakosság foglalkozásokhoz és nemekhez köthető attitűdjének méréséhez ezért saját kérdőívet készítettem Shinar (1975) tanulmánya alapján. A kérdőív a 3. Függelékben tekinthető meg.

Az Inspira Group kutatócég segítségével 1000 fős mintán (egy omnibusz vizsgálat során) lekérdezett kérdőívben arra voltam kíváncsi, hogy a megkérdezettek férfiasnak vagy nőiesnek tartanak-e egyes foglalkozásokat. A kérdőív elemzésénél a reprezentativitás érdekében a kutatócég saját súlyozását használtam. A kérdőívben összesen 100 foglalkozásról nyilatkoztak a megkérdezettek, ezen foglalkozások megtekinthetők a 2. Függelékben, a kiválasztásuk módjáról részletesebben *3.2. A foglalkozások kiválasztása* című fejezetben írok. Annak érdekében, hogy elkerüljük, hogy a válaszadók elfáradjanak a kérdőív kitöltése közben és ezzel növeljük a kérdőív megbízhatóságát, egy megkérdetettnek csak 20 foglalkozásról kellett eldöntenie, hogy férfiasnak vagy nőiesnek tartja-e. Ehhez a foglalkozásokat öt csoportra bontottam, a csoportokba a foglalkozások véletlenszerűen kerülhettek. Minden foglalkozást tartalmazó csoport 200-200 embertől lett lekérdezve. A lekérdezés során a 20 foglalkozás sorrendje a kérdőívben randomizálva volt annak érdekében, hogy csökkentsük a foglalkozások sorrendjéből adódó (a kifáradás miatt potenciálisan létrejövő) válaszadási torzítást.

A férfiasság-nőiesség mérésére egy 1-től 6-ig terjedő Likert-skálát használtam, ahol az 1-es jelentette azt, hogy a foglalkozást kifejezetten férfiasnak, a 6-os azt, hogy kifejezetten nőiesnek tartják. Ahhoz, hogy össze tudjam hasonlítani azt, hogy a Google Fordító milyen nemre fordította az adott foglalkozást azzal, hogy a megkérdezettek azt férfiasnak vagy nőiesnek tartják, a 6 fokú Likert-skálát két kategóriába kellett sűrítenem. A két kategória azt mérte, hogy milyen arányban tartják inkább férfiasnak és milyen arányban inkább nőiesnek az adott foglalkozást a megkérdezettek. Ennek kiszámításához figyelembe kellett venni azt, hogy nem ugyanannyira gondolják férfiasnak az adott foglalkozást például azok, akik az 1-es (kifejezetten férfias) és akik a 3-as (enyhén férfias) válaszlehetőséget adták – a nőiesnél ugyanígy. Ezért a válaszlehetőségeket egy, a Likert-skála eredeti skáláját figyelembe vevő súlyozással súlyoztam: az 1-es és 6-os (kifejezetten férfias és kifejezetten nőies)

válaszlehetőségeknek 2,5-szeres, a 2-es és 5-ös (közepesen férfias és közepesen nőies) válaszlehetőségeknek 1,5-szeres, a 3-as és 4-es (enyhén férfias és enyhén nőies) válaszlehetőségeknek 0,5-szeres súlyt adtam. Erre a súlyozásra látható példa a 2. és 3. Táblázatban**.**

A kérdőívben szerepelt egy kiegészítő kérdés is, ami azt hivatott felmérni, hogy mi alapján döntötték el a megkérdezettek, hogy férfiasnak vagy nőiesnek tartják az adott foglalkozást.

	1 – kifejezetten férfiasnak tartom	**2**	**3**	**4**	**5**	**6 – kifejezetten nőiesnek tartom**	**Összesen**
ács	170	12	7	3	4	0	200

2. Táblázat: A „Mennyire tartja férfiasnak vagy nőiesnek az alábbi foglalkozásokat egy 1-től 6-ig terjedő skálán? Az 1-es jelenti azt, hogy kifejezetten férfiasnak tartja, a 6-os jelenti azt, hogy kifejezetten nőiesnek. /ács" kérdésre érkezett válaszok gyakorisága a Likert-skálát figyelembe vevő súlyozás nélkül.

	1 – kifejezetten férfiasnak tartom	**2**	**3**	**4**	**5**	**6 – kifejezetten nőiesnek tartom**	**inkább férfias (%)**	**inkább nőies (%)**
ács	425	18	3,5	1,5	6	0	98%	2%

3. Táblázat: A „Mennyire tartja férfiasnak vagy nőiesnek az alábbi foglalkozásokat egy 1-től 6-ig terjedő skálán? Az 1-es jelenti azt, hogy kifejezetten férfiasnak tartja, a 6-os jelenti azt, hogy kifejezetten nőiesnek. /ács" kérdésre érkezett válaszok gyakorisága a Likert-skálát figyelembe vevő súlyozással, valamint annak megoszlása, hogy hány százalék tartja inkább férfiasnak és hány százalék tartja inkább nőiesnek a foglalkozást.

3.2. A foglalkozások kiválasztása

Prates és társai (2019) a Google Fordítót vizsgáló kutatásukhoz a foglalkozásokat az amerikai BLS 2017-es kérdőíves népességfelmérésében használt foglalkozási kategóriák alapján alakították ki. Ehhez hasonlóan a fordításokhoz szükséges foglalkozások kiválasztásához én a 2011-es Népszámlálásban használt FEOR kategóriákból és az azoknak *3.1.1. A*

foglalkoztatottak férfi-nő aránya című fejezetben leírt módon megfeleltetett SOC kategóriákból indultam ki. Ez elősegítette, hogy a kiválasztott foglalkozások fordításainak eredményeit össze lehessen vetni mind a magyar, mind pedig az amerikai foglalkozásstatisztikai adatokkal. Mivel a foglalkozások magyar-angol fordítását vizsgáltam, a foglalkozásokat magyarul határoztam meg. Emiatt elsősorban az eredetileg is magyar FEOR kategóriákból indultam ki. A FEOR kategóriák egy része megfeleltethető egy-egy konkrét foglalkozásnak (pl. 2241 – Állatorvos), de vannak olyan FEOR kategóriák is, amik nem csak egy foglalkozást fednek le, hanem többet is. Ez utóbbi kategóriák esetében szükség volt arra, hogy a kategóriát több foglalkozásra bontsam. Ebben segített az USA-ban használt, sokszor specifikusabb SOC rendszer. Amennyiben ezután sem volt egyértelmű, hogy a FEOR kategória milyen konkrét foglalkozásokat jelöl, a KSH által megjelölt, az egyes FEOR kódokhoz tartozó jellemző munkaköröket vettem alapul (Központi Statisztikai Hivatal é.n.[b]), de törekedtem arra, hogy az ismertebb – és így valószínűleg a Google Fordító korpuszában is többször előforduló – és ne a túl specifikus (például a csak magyarországi pozíciókat jelölő) foglalkozásokat válasszam ki. A FEOR kategóriákból származtatott foglalkozások definiálása után egy olyan táblázatot kaptam, melyben minden egyes foglalkozáshoz tartozik egy FEOR kategória és egy SOC kategória. Néhány eset volt, amikor a FEOR kategóriának hivatalosan megfelelő SOC kategória nem írta le pontosan az adott foglalkozást, ekkor azt egy jobban megfelelő SOC kategóriára javítottam a U.S. Department of Labor által fejlesztett foglalkozás kód kereső segítségével (National Center for O*NET Development 2020). Az előbb ismertetett szempontok szerint előállt foglalkozási listából mutatok be néhányat a 4. Táblázatban.

Bizonyos foglalkozásokat vagy foglalkozási kategóriákat kivettem a végső listából, mert (1) túl általánosak voltak vagy (2) mert önmagukban hordoztak nemre utaló jelölőket. Ez utóbbi kategóriába tartoztak az olyan foglalkozások, melyeknek csak női alakja van (pl. védőnő) vagy van külön női alakja (pl. színész-színésznő, titkár-titkárnő, tanár-tanárnő, szülész-szülésznő), hiszen ez befolyásolja, hogy a fordító „he”-re vagy „she”-re fordítja-e az adott foglalkozást. Amennyiben az ilyen foglalkozásoknak van semleges szinonimája vagy egy semleges megfogalmazása, úgy azt használtam a fordításoknál (pl. „tanár” helyett „pedagógus”-t, „középiskolai tanár”-t, „szülész” helyett „szülészorvos”-t). Ugyanezen okból nem szerepelnek a listán vallási foglalkozások, amelyek erősen nemhez köthetők. Az ezen szempontok figyelembevételével kapott listán 742 foglalkozás szerepel, a teljes listát az 1. Függelékben lehet megtekinteni.

foglalkozás	FEOR kategória	SOC kategória
állatorvos [a]	Állatorvos	Veterinarians
adatbázis-tervező [b]	Adatbázis-tervező és -üzemeltető	Database designers and administrators
adatbázis-üzemeltető [b]	Adatbázis-tervező és -üzemeltető	Database designers and administrators
fodrász [c]	Fodrász	Hairdressers, Hairstylists, and Cosmetologists
borbély [c]	Fodrász	Barbers

4. Táblázat: Néhány példa az előállt foglalkozási listából és a hozzájuk tartozó FEOR és SOC kategóriák.

a. A foglalkozási kategória egy konkrét foglalkozásnak feleltethető meg.

b. A FEOR kategóriába több foglalkozás tartozik.

c. A SOC kategóriába több foglalkozás tartozik.

A kérdőívben szereplő 100 foglalkozást is ebből a listából választottam ki. A kérdőívhez kiválasztott foglalkozásoknál a fő szempont az volt, hogy olyan foglalkozásokat válasszak ki, amelyeket teljesen lefed egy FEOR kategória annak érdekében, hogy a kérdőív eredménye, a népszámlálás eredménye és a fordítás eredménye teljes egészében összevethető legyen egymással. Emellett törekedtem arra, hogy a kérdőívben szereplő foglalkozások között vegyesen szerepeljenek olyan foglalkozások, melyeket főként férfiak és olyanok, melyeket főként nők töltenek be. Továbbá alacsony és magas státuszú foglalkozások is szerepeljenek benne. Ezen okból a kérdőívhez az eredeti foglalkozási listából 40-et azon foglalkozások közül választottam ki, amelyek szerepelnek egy foglalkozáspresztízs felmérésében (Giczi–Csányi 2018), a többi 60 foglalkozás a 2011-es népszámlálás szerinti legnépesebb foglalkozások közül kerültek ki. Ez utóbbi 60 foglalkozásnál is törekedtem arra, hogy egyensúlyban legyen a férfi és női munkavállalók aránya, illetve arra is törekedtem, hogy a foglalkozások lefedjenek különböző foglalkozáscsoportokat. Az elkészült lista a 2. Függelékben tekinthető meg.

3.3. A fordítások létrehozása

Ahhoz, hogy elemezni tudjam, hogy a Google Fordító az egyes foglalkozásokat hímnemre vagy nőnemre fordítja, a foglalkozásokat „ő…” kezdetű mondatokba kellett illesztenem. Fontos kiemelni, hogy a konkrét mondatokat többféleképpen is elő lehet állítani. A mondat kezdődhet kis és nagy „ő”-vel, lefordíthatjuk azt, hogy „ő orvos”, de azt is, hogy „ő egy orvos”. A mondatban ezek az apró változtatások is hatással lehetnek arra, hogy a fordító „he”-re vagy „she”-re fordítja az adott foglalkozást, ahogy azt az 5. Táblázat mutatja. Mivel szakdolgozatom

módszertana Prates és társai (2019) tanulmányára épül, akik „ő egy…” kezdetű mondatokat használtak a fordításhoz, amellett döntöttem, hogy esettanulmányomban én is „ő egy…” kezdetű mondatokat használjak.

Magyarul	Angolul
ő orvos	she’s a doctor
Ő orvos	He’s a doctor
ő egy orvos	he is a doctor

5. Táblázat: Mondat variációk az „orvos” szakmára és azok fordítása.

A foglalkozások „ő egy…” kezdetű mondatokba illesztését – a foglalkozások nagy számára való tekintettel – automatizáltam. Ehhez Python programban írt kódot használtam, amit a 4. Függelékbe illesztettem. Az így kapott mondatokat a Google Fordító dokumentumokat lefordító funkciójával fordíttattam le. A fordítások elvégzése után minden lefordított mondatról kiértékeltem, hogy női vagy férfi melléknevet használnak, a nemi torzítás mértékét ezek alapján számoltan ki.

3.4. A nemi torzítás mérése

A Google Fordítóban megjelenő nemi torzítás mérésére egy saját mérőszámot dolgoztam ki. A nemi torzítást és annak mértékét valamilyen elméleti alaphoz képest lehet mérni, egy olyan ideális fordítóhoz, ami nem torzít, ami a valós férfi-nő arányt tükrözi vissza. Ezért a torzítás méréséhez azt vizsgáltam, hogy a Google Fordító fordításai milyen mértékben térnek el az ideális fordító fordításaitól, az eltérésre pedig hibapontokat adtam. Az ideális, valós férfi-nő arányt visszatükröző fordító követheti a foglalkoztatottak férfi-nő arányát, vagy annak az arányát, hogy mennyien tartják férfiasnak és nőiesnek a foglalkozásokat. Mindkét esetben hasonló elven kellene működnie az ideális fordítónak, de az egyszerűség kedvéért először a foglalkoztatottak férfi-nő arányát megtartó ideális fordítót fejtem ki.

Tegyük fel, hogy „A” foglalkozást 60%-ban végzik nők és 40%-ban végzik férfiak. Egy olyan fordítónak, ami pontosan követi ezt az arányt, „A” foglalkozást az esetek 60%-ában kellene „she”-re és 40%-ában kellene „he”-re fordítania. Az ilyen működési elvű fordítót probabilisztikusnak nevezem, mert olyan valószínűséggel fordítja le az adott foglalkozást hímnemre és nőnemre, amilyen arányban a foglalkozást férfiak és nők végzik. A Google Fordító korábban, amikor minden mondatot csak egyféleképpen fordított le, nem így működött és jelenleg a dokumentumok fordítása esetén sem így működik. Minden egyes

nemsemleges mondat esetében döntenie kell arról, hogy azt angolul „he”-re vagy „she”-re fordítsa, és nem úgy működik, hogy egy foglalkozást az esetek 60%-ában „she”-re, 40%-ában pedig „he”-re fordítja le. Az ilyen típusú fordítót determinisztikus fordítónak nevezem. Egy ideális determinisztikus fordítónak „A” foglalkozást „she”-re kellene fordítania, ugyanis nagyobb arányban végzik nők, mint férfiak. Viszont ilyen esetben a fordító nem reprezentálja azt a 40% férfit, aki „A” foglalkozást végzi. Ez az ideális determinisztikus fordító hibája az ideális probabilisztikus fordítóval szemben, amit röviden ideális hibának nevezek (H_i). Az ideális hiba „A” foglalkozás esetében 40 hibapont. Ehhez a 40 pont ideális hibához kell hasonlítanunk a Google Fordító működését. Amennyiben a Google Fordító „A” foglalkozást „she”-re fordítja, akkor a fordító saját hibája (H_s) megegyezik az ideális determinisztikus fordító hibájával: 40 hibapont lesz. Ha viszont a Google Fordító „A” foglalkozást „he”-re fordítja, akkor nem reprezentálja egyáltalán azt a 60% nőt, aki a foglalkozást végzi, és így a fordító saját hibája 60 hibapont lesz. A Google Fordító torzítását (T) az ideális determinisztikus fordító hibájának (H_i) és a fordító saját hibájának (H_s) összehasonlításából számoltam ki az alábbi képlet alapján:

$$T = \frac{Hs - H_i}{H_i}$$

Abban az esetben, ha a Google Fordító jó nemre fordít – vagyis a Google Fordító saját hibája megegyezik az ideális determinisztikus fordító hibájával –, ezen számítás alapján a torzítás mértéke 0. Ha a Google Fordító nem jó nemre fordít, akkor a torzítás mértéke akármekkora pozitív szám lehet. Ha a Google Fordító „A” foglalkozást hímneműre fordítja, akkor a torzítás értéke 0,5, amit úgy értelmezhetünk, hogy az ideális determinisztikus fordítóhoz képest 50%-kal többet hibázik, 50%-kal jobban torzít.

A konkrét elemzés során a torzítás mértékét a magyar FEOR és az amerikai SOC foglalkozási kategóriák férfi-nő arányához hasonlítottam. Azt mértem, hogy a Google Fordító mennyire követi jól azt, hogy egy foglalkozási kategóriát több nő vagy több férfi végez. Azoknál a foglalkozási kategóriáknál, amikhez nem csak egy foglalkozás, hanem több foglalkozás tartozott, a torzítást a hozzá tartozó foglalkozásoknál lévő torzítás átlagaként számoltam ki. A 6. Táblázatban látható erre példa. A foglalkozásokkal kapcsolatos, kérdőívvel mért attitűdhöz képesti torzítás mérése is az előbb ismertetett képlet alapján történt. Amennyiben „A” foglalkozást a megkérdezettek 60%-a tartotta inkább nőiesnek és 40%-a inkább férfiasnak[6],

[6] A 100 kérdőívvel felmért foglalkozásnál ezek az arányok tartalmazzák *3.2. A foglalkozások kiválasztása* című fejezetben ismertetett súlyozásokat.

úgy az ideális determinisztikus fordító hibapontja 40 pont. Ha a Google Fordító rosszul, azaz hímneműre fordítja „A” foglalkozást, akkor a fordító saját hibapontja 60 pont, a torzítás értéke 0,5.

foglalkozás	FEOR kategória	a fordítás neme	a foglalkozást végző nők (%)	a foglalkozást végző férfiak (%)	torzítás a foglalkozásnál	torzítás a kategóriánál
állatorvosi asszisztens	Állatorvosi asszisztens	nő	38	62	0,6	0,6
textilműves	Textilműves, hímző, csipkeverő	férfi	81	19	3,3	1,1
hímző	Textilműves, hímző, csipkeverő	nő	81	19	0	
csipkeverő	Textilműves, hímző, csipkeverő	nő	81	19	0	

6. Táblázat: A torzítás kiszámítása egy olyan FEOR kategóriánál, ami egy foglalkozást takar és egy olyan FEOR kategóriánál, ami több foglalkozást takar.

A Google Fordító torzítását nem csak foglalkozási kategóriánként, hanem nagyobb foglalkozási csoportonként is elemeztem, mert kíváncsi voltam arra, hogy mennyire torzít a fordító például az egészségügyi szakmák esetében. Ehhez a FEOR és SOC kategóriákat nagyobb csoportokba rendeztem. A csoportok kialakításában figyelembe vettem a foglalkozások FEOR, SOC és ISCO struktúráját. Mind a három struktúrában eleve nagyobb csoportokba vannak rendezve az egyes foglalkozási kategóriák, de teljes mértékben nem követtem egyik struktúra csoportosítását sem a csoportok létrehozásakor. Olyan csoportok kialakítására törekedtem, amelyek tartalmukat tekintve hasonló foglalkozásokat tömörítenek, de mégsem túl tágak, mert az lehet, hogy elfedné a torzítást az egyes csoportokban. A kialakított 18 csoport a 7. Táblázatban látható. Mivel vannak olyan SOC kategóriák, amelyeknél a Bureau of Labour Statistics (2011) nem közölt adatokat a nők arányáról, ezért az egyes foglalkozási csoportokhoz tartozó foglalkozások és foglalkozási kategóriák eltérhetnek a magyar és az amerikai adatoknál. Azt, hogy mennyire torzít a Google Fordító az egyes foglalkozási csoportoknál, a csoportokba tartozó foglalkozási kategóriák torzításának súlyozott átlagával mértem. Így például az Egészségügyben mért torzításba nagyobb arányban számítottak azok az egészségügyi foglalkozások, amit sokan végeznek.

Foglalkozások	Magyar		Amerikai	
	foglalkozások száma	FEOR kategóriák száma	foglalkozások száma	SOC kategóriák száma
Vezetők	116	31	95	19
Műszaki, informatikai és természettudományi	98	69	65	28
Szociális, bölcsész- és társadalomtudományi	23	13	20	6
Egészségügyi	52	28	46	23
Kultúra, művészet, sport	71	27	41	13
Oktatás	23	14	22	8
Gazdasági	35	18	32	15
Jogi	5	5	12	3
Irodai, adminisztratív	36	27	32	20
Építőipari	29	20	24	14
Kézmű- és könnyűipari	33	23	13	6
Egyéb ipari	21	18	13	7
Szolgáltatás	77	46	72	37
Kereskedelem	21	14	17	13
Mezőgazdasági	20	15	17	4
Gépkezelői, szerelői	58	38	23	15
Járművezetői	22	15	11	8
Katonai	4	3	0	0

7. Táblázat: A foglalkozási csoportok és beletartozó foglalkozások és foglalkozási kategóriák száma a magyar FEOR és az amerikai SOC rendszer alapján.

3.5. Kiegészítő kutatás a melléknevekről

A foglalkozások mellett az elemzésembe mellékneveket is bevontam a Google Fordítóról készült esettanulmány kiegészítéseként. Ugyan a foglalkozások mellett a melléknevek fordításánál megjelenő nemi torzítást már vizsgálták kutatások a Google Fordító esetében, de ezek a kutatások önmagában vizsgálták a mellékneveket (Cho és társai 2019; Prates és társai 2019). Olyan mondatokat fordíttattak le a Google Fordítóval, mint „ő boldog”, „ő szomorú” (Prates és társai 2019). Én ezzel szemben a mellékneveket a foglalkozásokat tartalmazó mondatokba építettem, például olyan mondatok fordítását vizsgáltam, mint „ő egy jó orvos”. Tudomásom szerint ez az első olyan kutatás, ami a foglalkozások és a hozzá kapcsolt melléknevek fordításait vizsgálja a Google Fordító esetében. A fordításokhoz két melléknevet választottam ki és azokat fokoztam: jó, nagyon jó, rossz, nagyon rossz. A mondatok a következő mintát követték: „ő egy jó orvos”, „ő egy nagyon jó orvos”, „ő egy rossz orvos”, „ő

egy nagyon rossz orvos”. A mellékneveket tartalmazó mondatok előállításához ugyanazt a Python kódot használtam, mint az eredeti, csak foglalkozásokat tartalmazó mondatoknál, ami a 4. Függelékben található. A mondatok lefordításához pedig a Google Fordító dokumentumokat lefordító funkcióját használtam.

Mivel a jó és rossz szakemberek esetében nincs egy olyan összehasonlítási alap, mint ami önmagában a foglalkozások esetében a foglalkozásokat végző férfiak és nők aránya, vagy az, hogy mennyien tartják férfiasnak és mennyien nőiesnek az adott szakmát, ezért a mellékneveket is tartalmazó mondatoknál a torzítást nem tudtam olyan módszerrel mérni, ahogy azt a melléknév nélküli mondatok esetében tudtam. A melléknevet tartalmazó mondatok esetében azt vizsgáltam, hogy a melléknevek és azok fokozása hogyan változtatja meg a foglalkozások fordításának eredeti nemét.

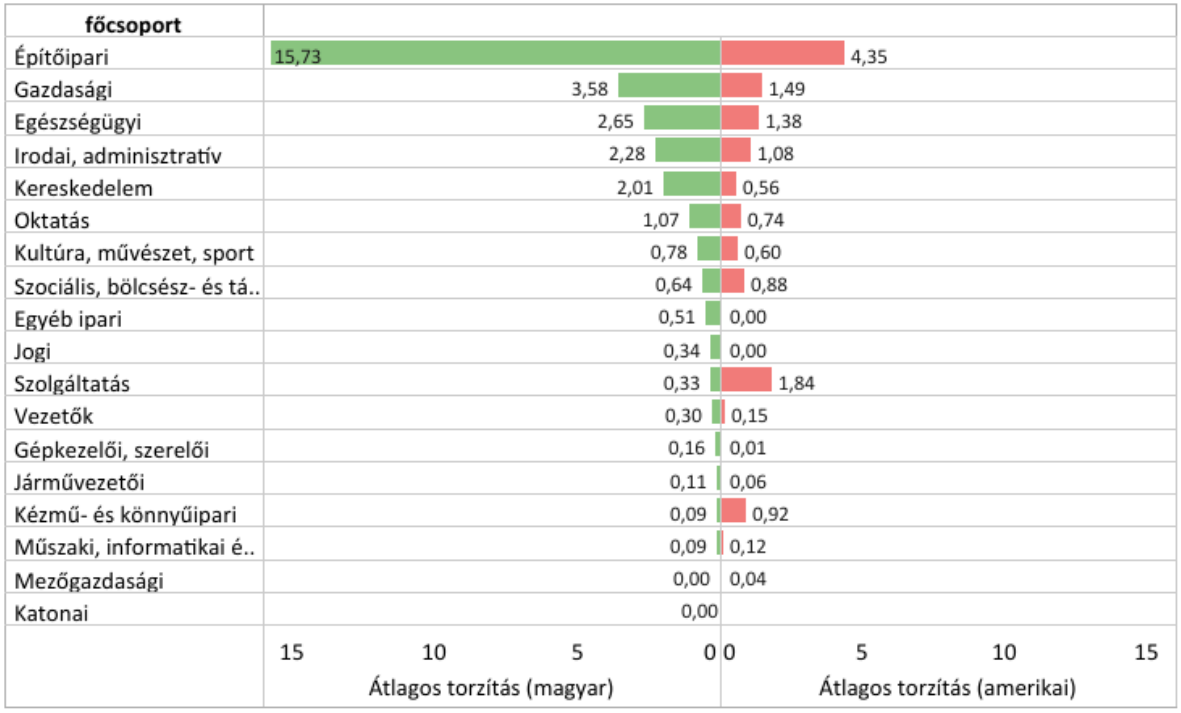
4. Elemzés

4.1. Torzítás a foglalkozások férfi-nő arányához képest

A Google Fordítóban mért nemi torzítás elemzésekor először a magyar foglalkozásstatisztikához mért torzítást fejtem ki. A magyar foglalkozási kategóriák 64%-át fordította jól és 36%-át fordította rosszul a Google Fordító. A rosszul fordított kategóriák 77%-át kellett volna nőnemű, 23%-át pedig hímnemű személyes névmással fordítania, vagyis a fordító többször torzított a nőkkel, mint a férfiakkal szemben. Azt is megvizsgáltam, hogy hány esetben torzított a fordító akkor, ha nőneműre és akkor, ha hímneműre kellett volna fordítania. Ha nőneműre kellett volna fordítania, az esetek 67%-ában hibázott, míg, ha hímneműre kellett volna fordítania, az esetek csupán 14%-ában hibázott. A fordítások amerikai foglalkozásstatisztikai adatokkal való összevetése hasonló mintázatot mutat. A Google Fordító a vizsgált SOC foglalkozási kategóriák 41%-ánál hibázott, ami egy kicsit több a magyar adatokból számolt hibázási aránynál. Ebből az esetek 73%-ában kellett volna nőneműre és 27%-ában hímneműre fordítania. A hibázás aránya az amerikai adatok esetében is a nőknél volt nagyobb: azon foglalkozások 70%-ánál fordított rosszul, amelyet nők végeznek többen, miközben azon foglalkozásoknak, amiket férfiak végeznek többen, a 20%-ánál fordított rosszul. Ezekből az adatokból leszűrhető, hogy a fordító nagyobb valószínűséggel torzít a nők esetében mind a magyar, mind az amerikai foglalkozásstatisztikai adatokhoz képest.

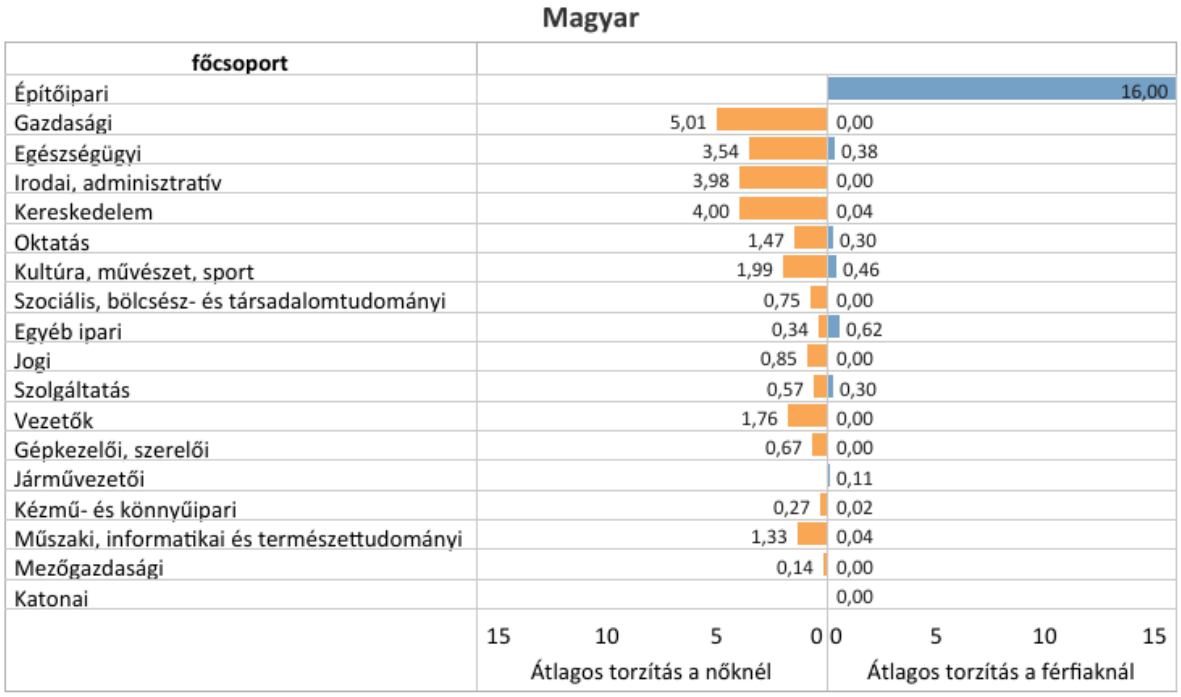
A következőkben a torzítás mértékének alakulását fejtem ki. A torzítási mérőszám 0-tól bármilyen nagy pozitív számig terjedhet, ahol a 0 jelenti azt, hogy nincs torzítás, a 0-ától eltérő torzítási érték pedig azt méri, hogy milyen mértékben torzít a Google Fordító egy ideális determinisztikus fordítóhoz képest. Ott, ahol van torzítás, a torzítás mértéke a magyar adatokhoz képest 0,001-től 173,44-ig, az amerikai adatokhoz képest 0,01-tól 88,91-ig terjed. Bár mindkét adathoz képest vannak ilyen nagy, kiugró torzítási értékek, a legtöbb foglalkozás esetében a torzítás mértéke ennél jóval alacsonyabb. A magyar adatoknál a medián torzítás 1,02, az amerikai adatoknál 0,92.

Azt, hogy milyen mértékben torzít a fordító az egyes foglalkozási főcsoportoknál, az 1. Diagram mutatja. A magyar adatoknál átlagosan az Építőipar területén a legnagyobb a torzítás, értéke jelentősen meghaladja a többi csoporthoz tartozó átlagos torzítást. Az Építőipart a Gazdaság és az Egészségügy követi. A Katonai foglalkozásoknál nem találtam torzítást. Az amerikai adatoknál szintén az Építőipar esetében a legmagasabb az átlagos torzítás, de kevésbé nagy, mint a magyar adatok esetében. Az amerikai adatokhoz mért torzítás átlagos mértéke az Építőipar után a Szolgáltatás és a Gazdaság esetében a legnagyobb. Az Egyéb ipari és Jogi foglalkozások esetében nem torzított a fordító.

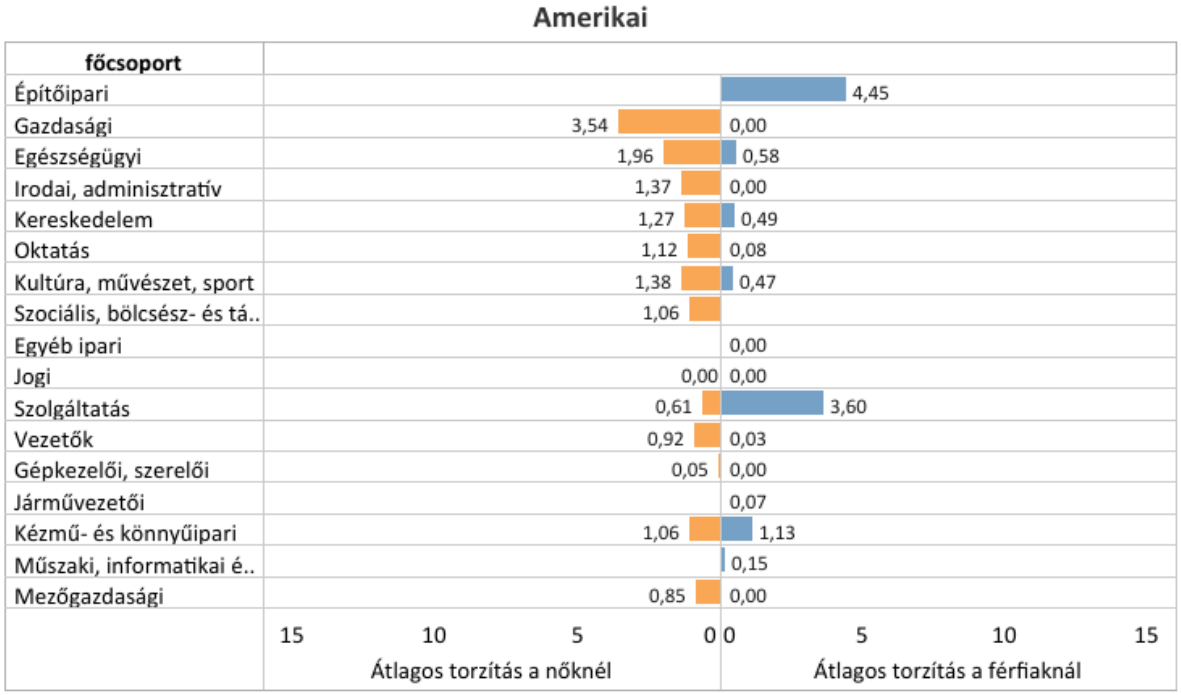


1. Diagram: A magyar és az amerikai adatokhoz mért torzítás átlaga az egyes foglalkozási főcsoportokban (súlyozva az egyes csoportokban lévő foglalkozásokat végzők számával).

Ezután azt vizsgáltam meg, hogy az egyes főcsoportokban a férfiakkal vagy a nőkkel szemben van-e torzítás. Ezt az indokolja, hogy vannak csoportok, amikben a férfi munkavállalók és vannak, amikben a női munkavállalók dominálnak. Ráadásul az egyes foglalkozási csoportokban vegyesen vannak olyan foglalkozások, amiket inkább nők és olyanok, amiket inkább férfiak végeznek. Ezért megvizsgáltam, hogy mekkora az átlagos torzítás a főcsoportokban, ha csak azokat a foglalkozásokat nézzük, amiket nőre kellene fordítania a Google Fordítónak és ha csak azokat, amiket férfire. A 2. Diagram a magyar adatokhoz mért átlagos torzításokat mutatja azoknál a foglalkozásoknál, amiket nőre és azoknál, amiket férfire kellene fordítania a Google Fordítónak. A 3. Diagram az amerikai adatokhoz mért átlagos torzításokat mutatja. A két diagramból jól látszik, hogy mely csoportokban és milyen mértékben torzít a fordító a nőkkel és a férfiakkal szemben.



2. Diagram: Átlagos torzítás főcsoportonként azoknál a foglalkozásoknál, amit nőre és amit férfire kellene fordítani a magyar foglalkozásstatisztika alapján.



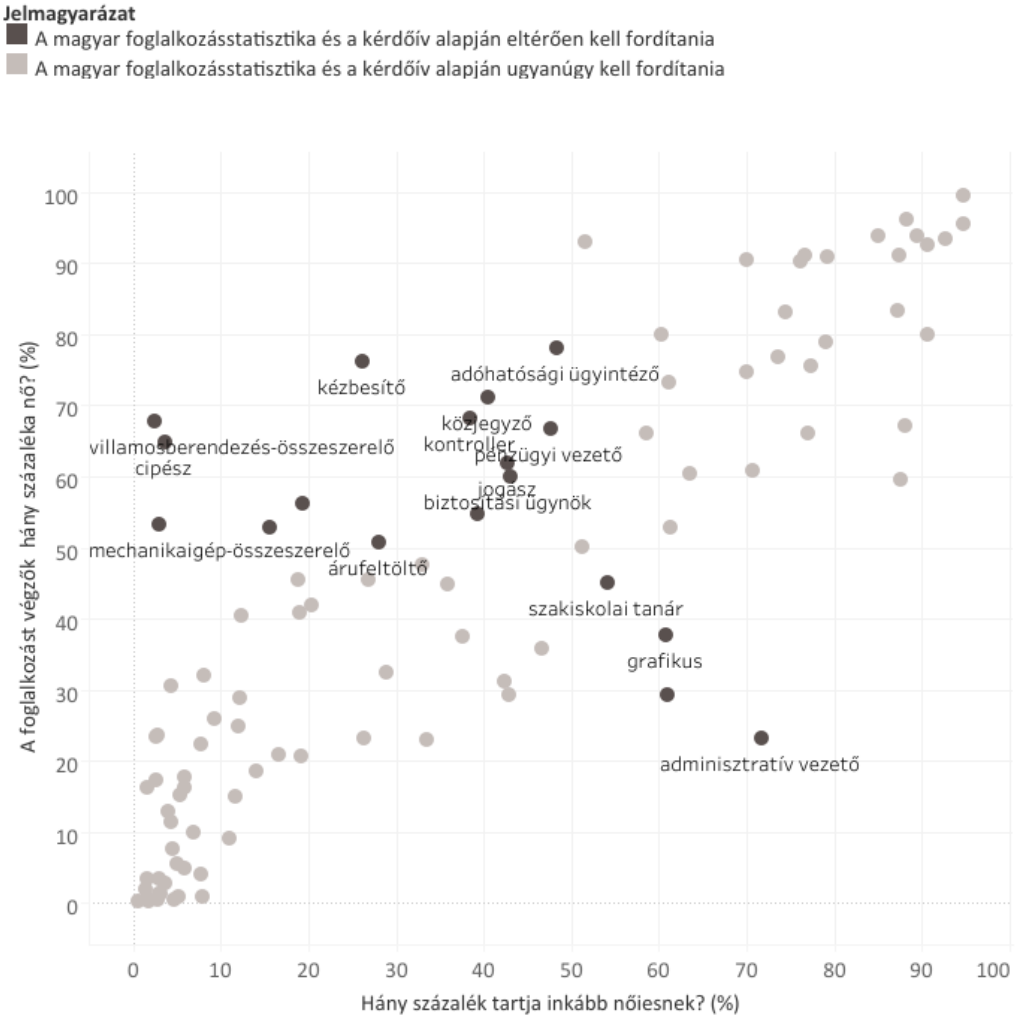
3. Diagram: Átlagos torzítás főcsoportonként azoknál a foglalkozásoknál, amit nőre és amit férfire kellene fordítani az amerikai foglalkozásstatisztika alapján.

Az Építőiparban, amelynél mindkét foglalkozásstatisztika alapján a legnagyobb volt az átlagos torzítás, a torzítást a férfiakkal szembeni torzítás okozza. Az amerikai adatoknál ezen kívül a Szolgáltatás területén volt jelentősebb a férfiakkal szembeni torzítás. Mindazonáltal a legtöbb főcsoport esetében a torzítás mértéke vagy a nőkkel szemben nagyobb, vagy a nőkkel és férfiakkal szemben közel azonos. Tehát ha főcsoportonként nézzük a torzítást, akkor elmondható, hogy néhány főcsoport kivételével a legtöbb főcsoportnál a torzítást a nőkkel szembeni torzítás okozza.

4.2. Torzítás a foglalkozásokkal kapcsolatos attitűdhöz képest

Ebben a fejezetben a foglalkozásokkal és a nemekkel kapcsolatos attitűdhöz mért torzítást vizsgálom, amit kérdőív segítségével térképeztem fel. A foglalkozások fordításait ahhoz hasonlítottam, hogy a megkérdezettek inkább nőiesnek vagy inkább férfiasnak tartják-e azokat. Annak eloszlását, hogy a megkérdezettek hány százaléka tartja inkább nőiesnek az egyes foglalkozásokat, a 4. Diagram mutatja. A diagramról az is leolvasható, hogy ez mennyire van összhangban azzal, hogy a foglalkozást végzők hány százaléka nő. A 45 fokos diagonális fölötti foglalkozásoknál nagyobb a nők aránya, mint azt az attitűd alapján várnánk. A 45 fokos diagonális alatti foglalkozásoknál pedig kisebb a nők aránya, mint ahogy azt az attitűd alapján

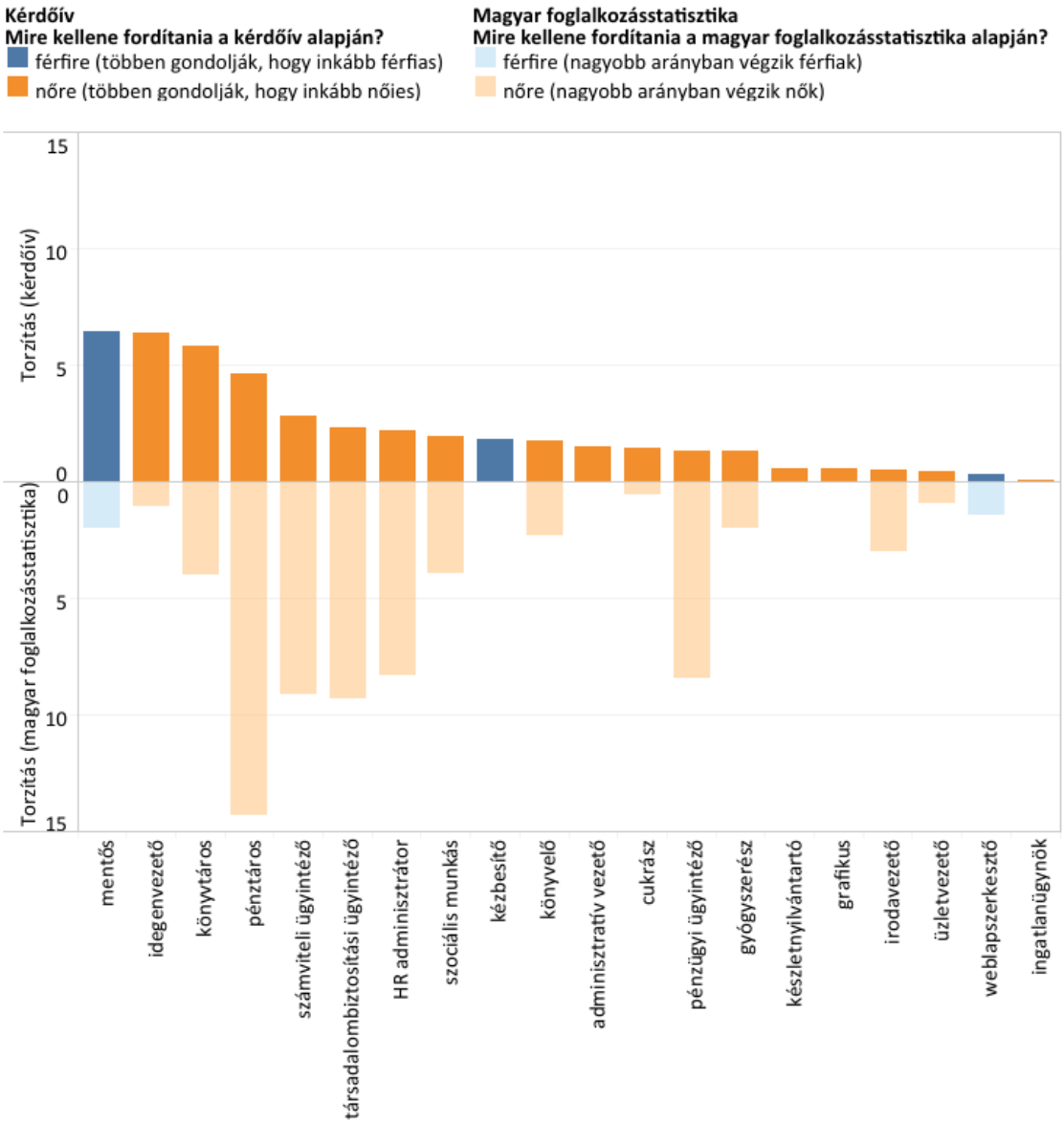
várnánk. Ezt lehet úgy interpretálni, hogy a társadalom foglalkozásokkal kapcsolatos attitűdje még nem követte le a társadalmi változást. A foglalkozások nagy részénél azonban az attitűd és a nők aránya összefügg egymással: a megkérdezettek azokat a foglalkozásokat tartották kevésbé nőiesnek, amit kevés nő végez és azokat tartották inkább nőiesnek, amit sok nő végez. Ezeknél a foglalkozásoknál a magyar foglalkozásstatisztika és a kérdőív alapján nézve is ugyanúgy kellene fordítania az algoritmusnak. Néhány kivétel esetében, mint „kézbesítő" vagy „adminisztratív vezető", a foglalkozást végző nők aránya és a foglalkozáshoz kapcsolódó attitűd nem függ össze egymással. Ezeknél a foglalkozásoknál egy ideális determinisztikus algoritmusnak máshogy kéne fordítania a magyar foglalkozásstatisztika és a kérdőív alapján.



4. Diagram: A foglalkozást végző nők aránya és a foglalkozáshoz kapcsolódó attitűd közötti kapcsolat.

A kérdőívben rákérdeztem arra, hogy mi alapján döntötték el a megkérdezettek azt, hogy egy foglalkozást férfiasnak vagy nőiesnek tartanak és a megkérdezettek 20%-a válaszolta

azt, hogy a foglalkozást végző férfiak és nők aránya alapján. 30%-uk döntött a foglalkozáshoz kapcsolódó személyiségjegyek alapján és 47%-uk a foglalkozáshoz szükséges fizikai képességek alapján. Voltak, akik több szempontot is figyelembe vettek és olyanok is, akik nehezen tudták eldönteni, mert szerintük „nincs már manapság olyan, hogy férfias, illetve nőies munkakör” (idézet a kérdőívre adott válaszokból).



5. Diagram: A kérdőív alapján rosszul fordított 20 foglalkozás torzításának mértéke a kérdőívhez és a magyar foglalkozásstatisztikához képest.

A következőkben a kérdőív alapján számolt torzítás mértékének elemzésére térek át. A kérdőívhez képest jobban teljesített a Google Fordító, mint a foglalkozásstatisztikákhoz képest. A 100 vizsgált foglalkozás 80%-át fordította úgy, ahogy azt egy ideális fordítónak

kellene, így az esetek egyötödében, 20 esetben hibázott. Összehasonlítva, a magyar foglalkozásstatisztika alapján számolva ebből a 100 foglalkozásból 30-nál fordított rosszul a program. A 20 esetből, amikor a kérdőívhez mérve hibázott az algoritmus, 17 esetben kellett volna nőneműre fordítania és csak 3 esetben hímneműre, ahogy az az 5. Diagramon látszik. A kérdőívhez mért torzítás mértéke az egyes foglalkozásoknál és annak összehasonlítása a magyar férfi-nő arányhoz mért torzítással szintén az 5. Diagramon látható.

Azt is megvizsgáltam, hogy az algoritmus milyen valószínűséggel torzít a nőkkel és a férfiakkal szemben a kérdőív eredményeihez képest. Amikor nőneműre kellett volna fordítania, az esetek 50%-ában torzított, amikor hímneműre, az esetek 4,5%-ában. Bár a 100 vizsgált foglalkozáshoz kötődő attitűdök nem reprezentálják a teljes foglalkozási listát, az arányok itt is azt a tendenciát tükrözik vissza, amit a foglalkozásstatisztikákhoz mért torzítások: a fordító a nőkkel szemben torzított gyakrabban.

4.3. Kiegészítő kutatás a melléknevekről

Melléknév	**A foglalkozások hány százalékát fordítja…**					
	nőre	**férfire**				
nincs	21%	79%				
jó	16%	84%				
nagyon jó	14%	86%				
rossz	10%	90%				
nagyon rossz	9%	91%				
Az eredeti mondatokhoz képest a foglalkozások hány százaléka…						
Melléknév	**maradt nő?**	**maradt férfi?**	**változott nőről férfire?**	**változott férfiról nőre?**	**nem változott**	**változott**
jó	16%	79%	5%	0%	95%	5%
nagyon jó	14%	79%	7%	0,1%	93%	7%
rossz	10%	79%	11%	0%	89%	11%
nagyon rossz	9%	79%	12%	0%	88%	12%

8. Táblázat: Hogyan változtatják meg a melléknevek azt, hogy milyen arányban fordította az algoritmus az eredeti, csak foglalkozásokat tartalmazó mondatokat nőneműre és hímenműre?

A foglalkozásokról készült esettanulmányt annak a vizsgálatával egészítettem ki, hogy hogyan változik meg a foglalkozások fordítása, ha a foglalkozások elé jelzőket teszünk: jó, nagyon jó, rossz, nagyon rossz. Önmagában mind a négy jelzőt hímneműre fordította az algoritmus, így a foglalkozások mellett mellékneveket tartalmazó mondatok fordításánál az volt várható, hogy ha változik a fordításban használt személyes névmás neme, inkább hímnemű személyes

névmásra változik, ami be is igazolódott. Azt, hogy ezek a melléknevek hogyan változtatták meg az eredeti, csak foglalkozásokat tartalmazó mondatok fordítását, a 8. Táblázat mutatja. Azt, hogy milyen nemmel fordítja az algoritmus a foglalkozásokat csak kis százalékban változtatták meg a jelzők. Ebből azt a konklúziót lehet levonni, hogy a vizsgált mondatokban a foglalkozásoknak nagyobb hatása van a fordításokra, mint a jelzőknek. Viszont érdemes megfigyelni, hogy a „rossz" és „nagyon rossz" jelzők körülbelül kétszer olyan sok foglalkozásnál változtatták meg a fordítás nemét, mint a „jó" és „nagyon jó" jelzők. Ahogy az várható volt, a jelzők egy eset kivételével hímneműre változtatták a személyes névmást.

5. Konklúzió

5.1. Összegzés

Szakdolgozatomban azzal a kérdéssel foglalkoztam, hogy a gépi tanuló algoritmusok, amelyek működését alapvetően racionálisnak, értékítéletektől mentesnek, objektívnek tartjuk, hogyan válhatnak igazságtalanná, előítéletessé egyénekkel vagy emberek egy csoportjával szemben. Mivel a gépi tanuláson alapuló algoritmusokat emberek írják, a működésükhöz és tanulásukhoz szükséges adatgyűjtést, adattisztítást emberek végzik, az emberi sztereotípiák, előítéletek, torzítások visszaköszönhetnek bennük, ami diszkriminatív hatással járhat egyénekre vagy emberek egy csoportjára nézve. Azt, hogy az algoritmikus torzítás jelensége pontosan hogyan alakul ki, ezt hogyan lehet vizsgálni és hogyan lehet kijavítani, a szakdolgozatomban a Google Fordítóban, egy gépi tanuláson alapuló online fordítóprogramban megjelenő nemi torzítás esetén keresztül mutattam be.

A Google Fordítóról készült esettanulmányban azt kutattam, hogy milyen mértékű az a nemi torzítás, ami foglalkozások magyar-angol fordításánál jelenik meg. A nemi torzítás mérésére egy saját mérőszámot dolgoztam ki, ami azt mérte, hogy a Google Fordító fordításai milyen mértékben térnek el egy ideálisan működő gépi fordító fordításaitól. Az ideális gépi fordítót három mutató alapján is meghatároztam: (1) a magyar lakosság férfi-nő aránya alapján az egyes foglalkozásokban, (2) az amerikai lakosság férfi-nő aránya alapján az egyes foglalkozásokban és (3) az alapján, hogy hányan és mennyire tartják az egyes foglalkozásokat inkább férfiasnak és inkább nőiesnek.

A Google Fordítóban megjelenő nemi torzítást vizsgáló korábbi kutatásokhoz hasonlóan azt találtam, hogy a fordító többször torzított a nőkkel, mint a férfiakkal szemben. A torzítás mértéke a legtöbb foglalkozási csoportban a nőkkel szemben volt nagyobb vagy a

nőkkel szemben közel azonos volt, mint a férfiakkal szemben. Korábbi kutatások azt mutatták, hogy a fordító gyakrabban fordít hímnemű személyes névmással, mint nőnemű személyes névmással és ezt a saját kutatásom is igazolta. Azoknál a mondatoknál, ahol a foglalkozásokhoz jelzőket kapcsoltam, ez a tendencia még erősebb volt. Ahol a jelzők megváltoztatták a fordítás nemét, ott a személyes névmás szinte mindig hímnemű személyes névmásra változott. Mindazonáltal az esetek kis százalékában történt csak változás a mellékneveknek köszönhetően, ami arra enged következtetni, hogy a vizsgált mondatokban a foglalkozásoknak nagyobb hatása van a fordításokra, mint a jelzőknek.

5.2. Az elemzés korlátai

Az esettanulmányban 742 foglalkozást vizsgáltam, ami nem fedi le az összes létező foglalkozást. Ennek ellenére, mivel a foglalkozásokat a magyar FEOR struktúra és az amerikai SOC struktúra alapján választottam ki, vélhetően a legnépesebb foglalkozások – amik valószínűsíthetően a Google Fordító korpuszában is gyakrabban fordulnak elő – bekerültek a foglalkozások listájába. Az amerikai adatokkal való összehasonlításba ennek a 742 foglalkozásnak csak egy részét tudtam bevonni, ugyanis némely SOC foglalkozási kategóriánál nem volt adatom a foglalkoztatottak férfi-nő arányával kapcsolatban. Ugyanezen okból az amerikai adatokkal való összehasonlításba a katonai foglalkozásokat sem tudtam bevonni.

A foglalkozásokhoz tartozó torzítás mértékének könnyebb bemutatása érdekében, illetve annak érdekében, hogy meg tudjam vizsgálni, hogy mely foglalkozási területeken nagyobb a torzítás, a foglalkozási kategóriákat foglalkozási csoportokra bontottam. Ezekbe a csoportokba a foglalkozásokat tematikájuk, tartalmuk szerint soroltam és például nem aszerint alakítottam ki a csoportokat, hogy egy csoportba kerüljenek a jellemzően nők által végzett foglalkozások és egy másikba a jellemzően férfiak által végzett foglalkozások. Emiatt a csoportokban vegyesen vannak inkább férfiak és inkább nők által végzett foglalkozások. Valamint a csoportokra vonatkozó férfi-nő arány eltér a csoportok között. Ezen kívül egyes csoportokban vegyesen lehetnek magasabb presztízsű és alacsonyabb presztízsű foglalkozások (pl. orvosok és ápolók). Ezek a különbségek befolyásolhatják az egyes csoportokban lévő átlagos torzítás mértékét. Ezt a problémát azzal próbáltam orvosolni, hogy minden csoportban külön megnéztem, hogy mely foglalkozásoknál nagyobb a torzítás: Azoknál, amiket nők végeznek többen (és így nőre kellene azokat fordítani) vagy azoknál, amiket férfiak végeznek többen (és így férfire kellene azokat fordítani)?

Annak eldöntéséhez, hogy egy ideális fordítónak milyen nemre kellene fordítania a mondatokat, kétféle indikátort vettem alapul: (1) a foglalkozásokat végző férfiak és nők arányát és (2) azt, hogy hányan és mennyire tartják inkább nőiesnek és inkább férfiasnak a foglalkozásokat. A kétféle indikátorral külön-külön hasonlítottam össze a fordításokat. Ez a gyakorlatban egy ideális fordító kidolgozásakor azért jelenthet problémát, mert bizonyos esetekben előfordult, hogy az egyik indikátor szerint nőre, a másik indikátor szerint férfire kellene fordítani a foglalkozást. Ezért a torzítás mérésére szükség lehet egy több indikátort kombináló mutató létrehozására.

5.3. További kutatási lehetőségek

A Google Fordítóról készült esettanulmány több ponton is további kutatási lehetőségeket rejt magában. A kutatást tovább lehet fejleszteni későbbi foglalkozásstatisztikai adatokkal való összevetéssel, annak tesztelésére, hogy a valóságban bekövetkező változások, mennyire jelennek meg a nyelvben és mennyire követi le azokat a Google Fordító algoritmusa. A torzítás mérését is tovább lehet fejleszteni például egy több indikátort tartalmazó komplex torzítási mutató kidolgozásával. Érdemes lehet megvizsgálni, hogy a foglalkozások presztízse és a foglalkozások fordításának neme milyen összefüggést mutat egymással: Jellemzően milyen nemre fordítja az algoritmus a magas és az alacsony presztízsű foglalkozásokat? A foglalkozásokon kívül más szöveg alapú torzításokat is lehet vizsgálni a fordítóprogramban, például melléknevek fordításait, ahogy azt Prates és társai (2019) valamint Cho és társai (2019) tették.

Elszakadva a gépi fordítástól, az algoritmikusok igazságos működésének vizsgálatában rengeteg kiaknázatlan lehetőség rejlik. Mivel a gépi tanulásban előforduló társadalmi torzításokkal csak nemrég kezdtek el foglalkozni kutatók, a témának sok olyan területe lehet, ami kevéssé vagy egyáltalán nem kutatott. Az algoritmikus torzítások és az algoritmikus diszkrimináció felderítése kiemelten fontos egy olyan világban, ahol számos üzleti, jogi, társadalmi döntés alapját képezik gépi tanuláson alapuló rendszerek. Érdemes tehát tovább kutatni a témában és tovább finomítani az ezek megtalálásához szükséges módszereket. A dolgozatban tárgyalt esettanulmány bepillantást engedett a problémába, annak vizsgálatába és kijavításának lehetőségébe.

Irodalomjegyzék

Angwin, Julia – Scheiber, Noam – Tobin, Ariana (2017): Facebook Job Ads Raise Concerns About Age Discrimination. *The New York Times*, december 20. (https://www.nytimes.com/2017/12/20/business/facebook-job-ads.html) (Utolsó megtekintés: 2020. március 16.)

Arzt, Samuel [Samuel Arzt] (2019 augusztus 23): *AI Learns to Park - Deep Reinforcement Learning*. [Videó] (https://www.youtube.com/watch?v=VMp6pq6_QjI) (Utolsó megtekintés: 2020. április 3.)

Beggs, Joyce M. – Doolittle, Dorothy C. (1993): Perceptions Now and Then of Occupational Sex Typing: A Replication of Shinar's 1975 Study. *Journal of Applied Social Psychology*, 23(17): 1435-1453.

Barocas, Solon – Selbst, Andrew D. (2016): Big Data's Disparate Impact. *California Law Review*, 104 (3): 671-732. http://dx.doi.org/10.15779/Z38BG31

Bolukbasi, Tolga – Chang, Ki-Wei – Zou, James – Saligrama, Venkatesh – Kalai, Adam (2016): Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee, Danial D. – von Luxburg, Ulrike – Garnett, Roman – Sugiyama, Masashi – Guyon, Isabelle (szerk.): *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*. NY, United States: Curran Associates Inc., 4349-4357.

Bureau of Labor Statistics (2011): Table 11. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity [Adattábla]. (https://www.bls.gov/cps/aa2011/cpsaat11.htm) (Utolsó megtekintés: 2020. április 15.)

Bureau of Labor Statistics (2015): Crosswalk between the 2008 International Standard Classification of Occupations to the 2010 SOC. [Táblázat] (https://www.bls.gov/soc/soccrosswalks.htm) (Utolsó megtekintés: 2020 április 15.)

Burrell, Jenna (2016): How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3 (1): 1-12. https://doi.org/10.1177/2053951715622512

Chen, Le – Ma, Ruijun – Hannák, Anikó – Wilson, Christo (2018): Investigating the Impact of Gender on Rank in Resume Search Engines. *Proceedings of the 2018 CHI Conference on*

Human Factors in Computing Systems, 651: 1–14. https://doi.org/10.1145/3173574.3174225

Cho, Won I. – Kim, Ji W. – Kim, Seok M. – Kim, Nam S. (2019): On Measuring Gender Bias in Translation of Gender-neutral Pronouns. *Association for Computational Linguistics,* Proceedings of the First Workshop on Gender Bias in Natural Language Processing: 173–181. https://doi.org/10.18653/v1/W19-3824

Couch, James V. – Sigler, Jennifer N. (2001). Gender Perception of Professional Occupations. *Psychological Reports*, 88(3): 693–698. https://doi.org/10.2466/pr0.2001.88.3.693

Cormen, Thomas H. – Leiserson, Charles E. – Rivest, Ronald L.– Stein, Clifford (2003): Algoritmusok. In uők.: *Új algoritmusok*. Budapest: Scolar Kiadó, 23-27.

Dastin, Jeffrey (2018): Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, október 10. (https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G) (Utolsó megtekintés: 2020. március 15.)

Freedman, David – Pisani, Robert – Purves, Roger (2005): Torzítások. In uők.: *Statisztika*. Budapest: Typotex Kiadó, 126-127.

Giczi Johanna – Csányi Gergely (2018): Mikrocenzus 2016 – 13. A foglalkozások presztízse. (http://www.ksh.hu/apps/shop.kiadvany?p_kiadvany_id=1040677) (Utolsó megtekintés: 2020. április 16.)

Goodfellow, Ian – Bengio, Yoshua – Courville, Aaron (2016): Machine Learning Basics. In uők.: *Deep Learning*. Cambridge, MA: MIT Press, 96-152. (http://www.deeplearningbook.org/) (Utolsó megtekintés: 2020. március 25.)

Goodman, Bryce – Flaxman, Seth (2016): European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38 (3): 50-57. https://doi.org/10.1609/aimag.v38i3.2741

[Google] (2017 augusztus 25): *Machine Learning and Human Bias*. [Videó] (https://www.youtube.com/watch?v=59bMh59JQDo&list=LLonNogqvSZZqOV3y0AUqTPA&index=2&t=0s) (Utolsó megtekintés: 2020 március 25)

Jurgens, David – Tsvetkov, Yulia – Jurafsky, Dan (2017): Incorporating Dialectal Variability for Socially Equitable Language Identification. *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics, 2: 51-57. https://doi.org/10.18653/v1/P17-2009

Kelman, Sveta (2014): Translate Community: Help us improve Google Translate! *Google Blog*, július 25. (https://search.googleblog.com/2014/07/translate-community-help-us-improve.html) (Utolsó megtekintés: 2020. március 25.)

Király Zoltán (2020): Algoritmus fogalma, modellezés. In uő.: *Algoritmuselmélet* (1.70 verzió). Typotex Kiadó, 9-10.

Központi Statisztikai Hivatal (2010): A FEOR-08 Négyszámjegyes rendszeres jegyzéke. Melléklet a 7/2010. (IV. 23.) KSH közleményhez. [Dokumentum] (https://www.ksh.hu/docs/szolgaltatasok/hun/feor08/pdf/feor08_kozl_melleklet.pdf) (Utolsó megtekintés: 2020. április 16.)

Központi Statisztikai Hivatal (é.n.[a]): A hazai (FEOR-08) és a nemzetközi (ISCO-08) foglalkozási nómenklatúrák közötti fordítókulcs. [Táblázat] (https://www.ksh.hu/docs/osztalyozasok/feor/fordkulcs_feor_isco_hu.pdf) (Utolsó megtekintés: 2020. április 15.)

Központi Statisztikai Hivatal (é.n.[b]): Foglalkozások Egységes Osztályozási Rendszere (FEOR–08). [Weblap] (https://www.ksh.hu/docs/szolgaltatasok/hun/feor08/feorlista.html) (Utolsó megtekintés: 2020. április 17.)

Kuczmarski, James (2018): Reducing gender bias in Google Translate. *Google Blog*, december 6. (https://www.blog.google/products/translate/reducing-gender-bias-google-translate/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+GoogleTranslateBlog+%28Translate+%7C+Google+Blog%29&utm_content=FeedBurner) (Utolsó megtekintés: 2020. március 25.)

Laki László János (2018): Mesterséges intelligencia a gépi fordításban. In: Tolcsvai Nagy Gábor (szerk.): *A humán tudományok és a gépi intelligencia*. Budapest: Gondolat Kiadó, 156-183.

Lovász László (2018): Számítási modellek. In uő.: *Algoritmusok Bonyolultsága* (1.4 verzió). 9-10.

Mitchell, Margaret [stanfordonline] (2019. április 5). *Stanford CS224N: NLP with Deep Learning | Winter 2019 | Lecture 19 – Bias in AI*. [Videó] (https://www.youtube.com/watch?v=XR8YSRcuVLE&list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&index=20&t=0s) (Utolsó megtekintés: 2020. március 25.)

Nagy Réka (2007): Új lencsék egy társadalmi jelenség vizsgálatában. A digitális egyenlőtlenségek kutatásának átfogó szemléletéről. *Szociológiai szemle*, 2007/1–2: 16–28.

National Center for O*NET Development (2020): Business and Financial Operations Occupations. *O*NET Code Connector*. [Weblap] (https://www.onetcodeconnector.org/find/family/title?s=13) (Utolsó megtekintés: 2020. április 28.)

Népszámlálás (2011): A foglalkoztatott népesség FEOR-08 kategóriák szerint. [Adattábla]

Olson, Parmy (2018): The Algorithm That Helped Google Translate Become Sexist. *Forbes*, február 15. (https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/#4c2537c17daa) (Utolsó megtekintés: 2020. március 25.)

Prates, Marcelo – Avelar, Pedro – Lamb, Luis C. (2019): Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 1-19.

Rangarajan Sridhar, Vivek Kumar (2015): Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words. In: Blunsom, Phil – Cohen, Shay – Dhillon, Paramveer – Liang, Percy (szerk.): *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Denver, Colorado: Association for Computational Linguistics, 192-200. https://doi.org/10.3115/v1/W15-1526

Sandvig, Christian – Hamilton, Kevin – Karahalios, Karrie – Langbort, Cedric (2014): *Auditing Algorithms. Research Methods for Detecting Discrimination on Internet Platforms*. (Az International Communication Association 64. éves konferenciáján tartott előadás, Seattle, WA, USA.)

Ságvári Bence (2017): Diszkrimináció, átláthatóság és ellenőrizhetőség. Bevezetés az algoritmikusetikába. *Replika*, 2017/3: 61-79.

Schiebinger, Londa (2014): Scientific research must take gender into account. *Nature*, 507: 9. https://doi.org/10.1038/507009a

Schwarm, Alex (2018): Amazon, Machine Learning, and Gender Bias. *LinkedIn*, október 31. (https://www.linkedin.com/pulse/amazon-machine-learning-gender-bias-alex-schwarm/) (Utolsó megtekintés: 2020. március 25.)

Shinar, Eva H. (1975): Sexual stereotypes of occupations. *Journal of Vocational Behavior*, 7 (1): 99-111.

Wu, Yonghui – Schuster, Mike – Chen, Zhifeng – Le, Quoc V. – Norouzi, Mohammad (2016): Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144. (http://arxiv.org/abs/1609.08144) (Utolsó megtekintés: 2020. március 25.)

Függelék

1. Függelék

A lefordított foglalkozások listája.

honvédtiszt
katonatiszt
honvéd tiszthelyettes
honvéd
alkotmánybíró
államtitkár
kormánybiztos
köztársasági elnök
legfőbb ügyész
miniszter
miniszterelnök
országgyűlési képviselő
önkormányzati képviselő
parlamenti képviselő
politikus
főügyész
főügyészhelyettes
bíróság elnöke
bíróság elnökhelyettese
bírósági kollégiumvezető
vezetőügyész
vezetőügyész-helyettes
rendőrfőkapitány
minisztériumi főosztályvezető
minisztériumi főosztályvezető-helyettes
minisztériumi osztályvezető
alpolgármester
polgármester
aljegyző
főjegyző
jegyző
körjegyző
polgármesteri hivatal osztályvezetője
önkormányzat osztályvezetője
önkormányzati tisztviselő
ágazati szakszervezet megyei vezetője
alapítványi elnök
alapítványi kuratóriumi elnök
alapítványi ügyvezető igazgató
ipartestületi igazgató
közalapítványi vezető
érdekvédő szervezet vezetője
pártelnök
pártfrakció-vezető
szakszervezeti elnök
szakszervezeti vezető
vezérigazgató
ügyvezető igazgató
rektor
rektorhelyettes
ipari kisszövetkezet elnöke
kórházigazgató
főorvos
múzeumigazgató
mezőgazdasági egység vezetője
erdészetvezető
halászati egység vezetője
vadászati egység vezetője
asztalosműhely-vezető
fűrészüzemvezető
gyáregység-vezető
gyárrészlegvezető
ipari szervezet vezetője
ipariüzem-vezető
ipartelep-vezető
műszaki igazgató
műszaki igazgatóhelyettes
termelési igazgató
gyárigazgató
építésvezető
szállítási vezető
logisztikai vezető
raktározási egység vezetője
informatikai tevékenységet folytató egység vezetője
telekommunikációs tevékenységet folytató egység vezetője
biztonsági szolgálat vezetője
börtönigazgató
javítóintézeti igazgató
magánnyomozó-iroda vezetője
őrző-védő szolgálat vezetője
rendészeti vezető
rendőrfelügyelő
bankelnök
szociális tevékenységet folytató egység vezetője
óvoda igazgató
idősgondozási tevékenységet folytató egység vezetője
egészségügyi tevékenységet folytató egység vezetője
iskolaigazgató
dékán
dékánhelyettes
tanszékvezető
szállodaigazgató
étteremvezető
kereskedelmi vezető
boltvezető
élelmiszerboltos
üzleti szolgáltatási tevékenységet folytató egység vezetője
art direktor
balettigazgató
könyvtárigazgató
levéltári igazgató
művelődési ház igazgatója
művészeti galéria vezetője
művészeti vezető
színházigazgató
kaszinóvezető
sportegyesület vezetője
uszodavezető
pénzügyi vezető
hr vezető
személyzeti vezető
kutatási és fejlesztési tevékenységet folytató egység vezetője
stratégiai vezető
marketingvezető
marketingmenedzser
értékesítési menedzser
reklám-vezető

pr-vezető
kommunikációs vezető
adminisztratív vezető
bányamérnök
kohó- és anyagmérnök
élelmiszer-ipari mérnök
faipari mérnök
könnyűipari mérnök
építészmérnök
építőmérnök
vegyészmérnök
gépészmérnök
villamosmérnök
energetikai mérnök
elektronikai mérnök
telekommunikációs mérnök
mezőgazdasági mérnök
természetvédelmi mérnök
erdőmérnök
tájépítész
kertépítő mérnök
várostervező
földmérő
térinformatikus
térképész
grafikus
minőségbiztosítási mérnök
informatikai rendszerelemző
szoftverfejlesztő
informatikus
hálózat- és multimédia-fejlesztő
alkalmazásfejlesztő
programozó
weblapszerkesztő
adatbázis-üzemeltető
adatbázis-tervező
rendszergazda
számítógép-hálózati üzemeltető
informatikai biztonság szakértő
fizikus
csillagász
meteorológus
kémikus
geológus
matematikus
zoológus
botanikus
biokémikus
biofizikus
mikrobiológus
biológus
környezetvédelmi mérnök
háziorvos
körzeti orvos
gyermekorvos
orvos
nőgyógyász
aneszteziológus
kardiológus
bőrgyógyász
sebész
plasztikai sebész
neurológus
patológus
radiológus
fogorvos
szájsebész
gyógyszerész
higiénia ellenőr
járványügyi felügyelő
optometrista
dietetikus
fizioterapeuta
gyógytornász
mentős
hallás- és beszédterapeuta
természetgyógyász
egészségfejlesztő
egészségügyi referens
egészségügyi szaktanácsadó
művészetterapeuta
szülészorvos
állatorvos
növényorvos
szociálpolitikus
szociális munkás
egyetemi oktató
főiskolai oktató
középiskolai tanár
pedagógus
szakiskolai tanár
általános iskolai tanár
óvodapedagógus
gyógypedagógus
konduktor
fejlesztő tanár
tanfelügyelő
nyelvtanár
zenetanár
drámapedagógus
művészitorna-oktató
néptáncoktató
tánctanár
hitoktató
nevelőtiszt
kollégiumi nevelőtanár
könyvtár pedagógus
pénzügyi elemző
befektetési tanácsadó
adószakértő
könyvelő
kontroller
vezetési tanácsadó
üzletpolitikai elemző
karrier-tanácsadó
személyzeti szakember
továbbképzési és személyzetfejlesztési szakértő
piackutató
marketinges
kommunikációs elemző
médiatanácsadó
pr felelős
szóvivő
értékesítési tanácsadó
exportfelelős
importfelelős
kereskedelmi szervező
üzletlánc felelős
vevőköri felelős
jogász
ügyész
közjegyző
ügyvéd
politológus
filozófus
történész
régész
néprajzkutató
közgazdász
statisztikus
szociológus
demográfus
nyelvész
fordító
tolmács
klinikai szakpszichológus
iskolapszichológus
pszichológus
munkapszichológus
grafológus

kriminológus
antropológus
archeológus
geográfus
könyvtáros
levéltáros
muzeológus
kurátor
kulturális szervező
népművelő
producer
produkciós menedzser
programigazgató
programszervező
fesztiválszervező
koncertszervező
szerkesztő
főszerkesztő
újságíró
riporter
rádióműsor-szerkesztő
televízióműsor-szerkesztő
filmkritikus
külföldi tudósító
tudósító
műsorvezető
sportriporter
edző
képzőművész
festőművész
szobrász
illusztrátor
iparművész
ruhatervező
zeneszerző
zenész
rendező
operatőr
humorista
bábművész
táncművész
koreográfus
akrobata
állatidomár
artista
bohóc
bűvész
kötéltáncos
légtornász
pantomimművész
zsonglőr
bányászati technikus
kohó- és anyagtechnikus
élelmiszer-ipari technikus
faipari technikus
könnyűipari technikus
vegyésztechnikus
gépésztechnikus
építő- és építésztechnikus
villamosipari technikus
mezőgazdasági technikus
erdővédelmi technikus
természetvédelmi technikus
térinformatikai technikus
földmérő technikus
környezetvédelmi technikus
minőségbiztosítási technikus
műszaki rajzoló
építőanyag-ipari technikus
gépjármű-üzemeltetési technikus
sugárzásmérő
közlekedési technikus
informatikai és kommunikációs rendszereket kezelő technikus
helpdesk operátor
rendszeradminisztrátor
számítógépes rendszerkarbantartó
számítógépes hálózati technikus
webtechnikus
műsorszóró és audiovizuális technikus
telekommunikációs technikus
erőműkezelő
vízművi berendezés kezelő
égetőművi berendezés kezelő
csatornaművi berendezés kezelő
vegyipari feldolgozó berendezés kezelő
kőolaj- és földgázfinomító berendezés kezelő
fémgyártási berendezés kezelő
ipari és termelési mérnök
szállítmányozási nyilvántartó
energetikus
munkavédelmi felügyelő
tűzrendész
hajóparancsnok
fedélzeti tiszt
hajózómérnök
pilóta
légiforgalmi irányító
légiforgalmi biztonsági technikus
légiközlekedési technikus
bányaműszak-vezető
főaknász
irodavezető
séf
konyhafőnök
ápoló
szakápoló
szülészeti asszisztens
orvosi asszisztens
orvosírnok
képi diagnosztikai asszisztens
sugárterápiás szakasszisztens
röntgenasszisztens
orvosi és laboratóriumi technikus
dentálhigiénikus
fogászati asszisztens
gyógyszertári asszisztens
fizioterápiás asszisztens
masszőr
fogtechnikus
ortopédiai eszközkészítő
látszerész
szemész
állatorvosi asszisztens
oktatási asszisztens
szociális segítő
gyermekgondozó
szociális gondozó
jelnyelvi tolmács
ifjúságsegítő
munka-közvetítő
pénzügyi ügyintéző
hitelügyintéző
banki ügyintéző
kölcsönügyintéző
pénzintézeti ügyintéző
bróker
tőzsdeügynök

számviteli ügyintéző
statisztikai ügyintéző
kárszakértő
kárbecslő
értékbecslő
biztosítási ügynök
kereskedelmi ügyintéző
anyaggazdálkodó
kereskedelmi üzletkötő
rendezvényszervező
esküvőszervező
marketing- és pr-ügyintéző
ingatlanügynök
személyi asszisztens
jogi asszisztens
vám- és pénzügyőr
adóhatósági ügyintéző
társadalombiztosítási ügyintéző
okmányirodai ügyintéző
nyomozó
végrehajtó
adósságbehajtó
anyakönyvvezető
hagyatéki ügyintéző
tűzvédelmi ellenőr
vadőr
statiszta
segédrendező
fényképész
díszlettervező
jelmeztervező
lakberendező
állatpreparátor
restaurátor
múzeumi technikus
régészeti ásatási technikus
könyvtári technikus
disc-jockey
dj
lemezlovas
tv bemondó
hírolvasó
öltöztető
sportoló
fitnesz edző
irodai adminisztrátor
gépíró
adatrögzítő
kódoló
bérelszámoló
pénzügyi adminisztrátor
statisztikai adminisztrátor
biztosítási adminisztrátor
készletnyilvántartó
könyvtári nyilvántartó
levéltári nyilvántartó
hr adminisztrátor
postás
iratkezelő
irattáros
banki pénztáros
játéktermi felügyelő
játékgépkezelő
játéktermi osztó
krupié
bukméker
zálogházi ügyintéző
pénzkölcsönző
utazási ügynök
utazási irodai ügyintéző
recepciós
szállodai recepciós
telefonos ügyfélszolgálati ügyintéző
ügyfélszolgálati ügyintéző
lakossági kérdező
régiségkereskedő
kereskedő
autókereskedő
könyvkereskedő
műkereskedő
virágkereskedő
üzletvezető
bolti eladó
trafikos
dohányboltos
piaci árus
pénztáros
benzinkutas
modell
telefonos értékesítési ügynök
házaló ügynök
vendéglős
kocsmáros
büfés
felszolgáló
pultos
csapos
cukrász
borbély
fodrász
sminkes
kozmetikus
manikűrös
pedikűrös
műkörmös
asztrológus
segédápoló
műtőssegéd
házi gondozó
kalauz
menetjegyellenőr
utaskísérő
légiutas-kísérő
idegenvezető
takarító szolgálatvezető
gondnok
rendőr
tűzoltó
börtönőr
vagyonőr
testőr
természetvédelmi őr
közterület-felügyelő
vízimentő
alpinista
járművezető-oktató
hobbiállat-gondozó
hobbiállat-kozmetikus
temetkezési vállalkozó
gyepmester
úszómester
növénytermesztő
zöldségtermesztő
gyümölcstermesztő
szőlész
kertész
gyógynövénytermesztő
állattenyésztő
baromfitenyésztő
méhész
kutyatenyésztő
állatgondozó
vegyes gazdálkodó
őstermelő
gazda
erdész
favágó
fakitermelő
vadász
halász
hentes
gyümölcs- és zöldségfeldolgozó

tejfeldolgozó
pék
borász
szabásminta-készítő
szabó
kalapos
kesztyűs
szűcs
szőrmefestő
tímár
bőrdíszműves
cipész
famegmunkáló
esztergályos
bútorasztalos
asztalos
kárpitos
kádár
bognár
nyomdai előkészítő
nyomdász
könyvkötő
fémöntőminta-készítő
lakatos
szerszámkészítő
forgácsoló
csiszoló
köszörűs
hegesztő
lángvágó
kovács
fényező
árbocmester
gépjárműkarbantartó
autószerelő
motorkarbantartó
légijármű szerelő
mezőgazdasági gép szerelő
iparigép szerelő
műszerész
kerékpárszerelő
elektroműszerész
informatikai berendezések műszerésze
telekommunikációs berendezések műszerésze
elektromoshálózat-szerelő
címfestő
ékszerész
ékszerkészítő
ötvös
drágakőcsiszoló
keramikus
fazekas
üvegfújó
hangszerkészítő
szőr- és tollfeldolgozó
textilműves
hímző
csipkeverő
órás
kőműves
gipszkartonozó
stukkoló
ács
épületasztalos
betonozó
vízvezeték szerelő
gázszerelő
klímagépszerelő
légkondicionáló-szerelő
hűtőberendezés-szerelő
felvonószerelő
villanyszerelő
szigetelő
tetőfedő
fémlemez-megmunkáló
burkoló
szobafestő
mázoló
tapétázó
kőfaragó
műköves
kályha- és kandallóépítő
üveges
búvár
ipari alpinista
robbantómester
kártevőírtó
kéményseprő
aszfaltozó
útépítő
csatornafektető
pályamunkás
vágányfektető
élelmiszergyártó gép kezelő
italgyártó gép kezelő
dohánygyártó gép kezelő
tisztítógép-kezelő
ruhafestőgép-kezelő
fehérítőgép-kezelő
szövő- és kötőgépkezelő
kötőgépkezelők
fonalfonó gép kezelő
fonalelőkészítő gép kezelő
fonalsodró gép kezelő
szőrme- és bőr-kikészítőgép kezelő
cipőgyártó gép kezelő
fafeldolgozó gépkezelő
papírtermék gyártó gépkezelő
kőolaj- és földgázfeldolgozó gép kezelő
vegyi alapanyagot gyártó gép kezelő
gyógyszergyártó gép kezelő
műtrágya- és növényvédőszer-gyártó gép kezelő
műanyagtermék-gyártó gépkezelő
gumitermék gyártó gépkezelő
fotó- és mozgófilmlaboráns
kerámiaipari terméket gyártó gép kezelő
üvegterméket gyártó gép kezelő
cement-feldolgozó gép kezelő
kőfeldolgozó gép kezelő
ásványianyag-feldolgozó gép kezelő
papírgyártó berendezés kezelő
fémfeldolgozó gépkezelő
fémmegmunkáló gép kezelő
felületkezelő gép kezelő
mechanikaigép-összeszerelő
villamosberendezés-összeszerelő
kőfejtő
bányász
kútfúró
kazángépkezelő
csomagológép kezelő
palackozógép kezelő
címkézőgép kezelő
mozigépész
mosodai gép kezelő
vonatvezető
bakter
vasutas
villamosvezető
metróvezető

trolibuszvezető
sofőr
taxisofőr
kamionsofőr
tehergépkocsi-vezető
traktoros
buszvezető
mezőgazdasági gép kezelő
erdőgazdasági gép kezelő
növényvédő gép kezelő
földmunkagép kezelő
köztisztasági gép kezelő
darukezelő
felvonógép kezelő
targoncavezető
matróz

mosodai dolgozó
autómosó
kocsimosó
gépkocsimosó
ablaktisztító
utcaseprő
szemétszállító
hulladékválogató
állati erővel vont jármű hajtója
rakodómunkás
árufeltöltő
csomagoló
biztonsági őr
portás
telepőr

mérőóra-leolvasó
kézbesítő
hordár
csomagkihordó
futár
pizzafutár
gyorséttermi eladó
konyhai kisegítő
mosogató
parkolóőr
bányászati segédmunkás
kőfejtő segédmunkás
kubikos
építőipari segédmunkás
mezőgazdasági munkás

2. Függelék

A kérdőívhez kiválasztott 100 foglalkozás 5 csoportra bontva. Egy-egy csoport foglalkozásait 200 megkérdezett értékelte abból a szempontból, hogy mennyire férfiasak vagy nőiesek.

1	2	3
ács	adóhatósági ügyintéző	építőipari segédmunkás
szociális munkás	elektroműszerész	műanyagtermék-gyártó gépkezelő
építésvezető	közgazdász	biztosítási ügynök
műszerész	gyógypedagógus	rendszergazda
rakodómunkás	orvosi asszisztens	ügyvéd
villanyszerelő	lakatos	állatorvos
grafikus	irodavezető	óvodapedagógus
állattenyésztő	pénzügyi vezető	pénzügyi ügyintéző
forgácsoló	matematikus	pék
mezőgazdasági munkás	nyomdász	gyógyszerész
mentős	cipész	kőműves
szakiskolai tanár	gépészmérnök	közjegyző
könyvelő	adminisztratív vezető	könyvtáros
irodai adminisztrátor	építészmérnök	kontroller
ingatlanügynök	benzinkutas	targoncavezető
mechanikaigép-összeszerelő	nyelvtanár	növénytermesztő
vegyészmérnök	étteremvezető	felszolgáló
kertész	jogász	órás
elektronikai mérnök	borász	HR adminisztrátor
készletnyilvántartó	üzletpolitikai elemző	cukrász

4	5
villamosberendezés-összeszerelő	tűzoltó
hentes	házi gondozó
rendőr	anyaggazdálkodó
weblapszerkesztő	társadalombiztosítási ügyintéző
bankelnök	mozdonyvezető
idegenvezető	gépésztechnikus
buszvezető	honvéd tiszthelyettes
üzletvezető	gyógyszertári asszisztens
személyi asszisztens	szerszámkészítő
informatikai rendszerelemző	lakberendező
postás	parkolóőr
minőségbiztosítási technikus	számviteli ügyintéző
csomagoló	szövő- és kötőgépkezelő
kereskedelmi ügyintéző	földmunkagép kezelő
gondnok	pénztáros
recepciós	fémfeldolgozó gépkezelő
árufeltöltő	burkoló
kézbesítő	szabó
fényképész	szociális gondozó
fafeldolgozó gépkezelő	esztergályos

3. Függelék

A kérdőív kérdései egy minta foglalkozással.

1) Mennyire tartja férfiasnak vagy nőiesnek az alábbi foglalkozásokat egy 1-től 6-ig terjedő skálán? Az 1-es jelenti azt, hogy kifejezetten férfiasnak tartja, a 6-os jelenti azt, hogy kifejezetten nőiesnek.

		1 - kifejezetten férfiasnak tartom	**2**	**3**	**4**	**5**	**6 - kifejezetten nőiesnek tartom**
1	ács						

2) Milyen szempont alapján döntötte el, hogy az adott foglalkozás férfias vagy nőies? Kérem, az alábbi szempontok közül egyet jelöljön meg!

1 – A foglalkozást végző férfiak és nők aránya alapján.

2 – A foglalkozáshoz kapcsolódó személyiségjegyek alapján.

3 – A foglalkozáshoz szükséges fizikai képességek alapján.

4 – Egyéb, éspedig:

4. Függelék

A foglalkozásokat és a kiegészítő kutatáshoz a mellékneveket tartalmazó mondatok előállításához használt Python kód.

```
melleknev_dict = {None : '',
                  'jó' : 'jo',
                  'nagyon jó' : 'nagyon_jo',
                  'rossz' : 'rossz',
                  'nagyon rossz' : 'nagyon_rossz'}
```

```
def save_sentence(infajlhely='foglalkozasok.txt', outfajlhely='./', mondat='ő egy', melleknev=None):

    """
    Beolvassa a foglalkozásokat, mondatba foglalja azokat és kiírja egy txt fájlba a mondatokat.

    :param infajlhely: a foglalkozásokat tartalmazó fájl helye (string)
    :param mondat: a mondat eleje (string), a default értéke 'ő  egy' (lehetne még: 'ő', 'Ő')
    :param melleknev: a mondathoz adott melléknév (string), pl. 'jó ', alapbeállításban nincs megadva
    :param outfajlhely: a txt fájl helye (string)
    """
    #A foglalkozások beolvasása.
    with open(infajlhely, encoding='utf-8') as infile:
        foglalkozasok = infile.read().lower().split('\n')

    #A mondatok elmentése
    outfajlnev = f'{outfajlhely}mondat_{melleknev_dict[melleknev]}.txt'

    with open(outfajlnev, 'w') as outfile:
        for foglalkozas in foglalkozasok:

            #Foglalkozások mondatba foglalása.
            if melleknev is None:
                teljes_mondat = f"{mondat} {foglalkozas}"
            else:
                teljes_mondat = f"{mondat} {melleknev} {foglalkozas}"

            outfile.write(teljes_mondat+'\n')
```

```
for opcio in melleknev_dict.keys():
    save_sentence(melleknev=opcio)
```